1-1-2018

Computational Analysis of MAP3K Kinases across Plant Genomes and Functional Characterization of a Subset of MAP3Ks in Nematode Resistance

Nobert Tamas Bokros

Computational analysis of MAP3K kinases across plant genomes and functional characterization of a subset of MAP3Ks in nematode resistance

By

Norbert Tamas Bokros

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Biochemistry
in the Department of Biochemistry, Molecular Biology, Entomology, & Plant Pathology

Mississippi State, Mississippi

December 2018

Computational analysis of MAP3K kinases across plant genomes and functional characterization of a subset of MAP3Ks in nematode resistance

By

Norbert Tamas Bokros

Approved:

_____

Sorina C. Popescu
(Major Professor)

_____

George V. Popescu
(Committee Member)

_____

Martin Wubben
(Committee Member)

_____

Kenneth O. Willeford
(Graduate Coordinator)

_____

George Hopper
Dean
College of Agriculture and Life Sciences

Name: Norbert Tamas Bokros

Date of Degree: December 14, 2018

Institution: Mississippi State University

Major Field: Biochemistry

Major Professor: Sorina C. Popescu

Title of Study: Computational analysis of MAP3K kinases across plant genomes and functional characterization of a subset of MAP3Ks in nematode resistance

Pages in Study 115

Candidate for Degree of Master of Science

Plant MAP3Ks have expanded significantly compared to their metazoan counterparts. A new, sequential workflow combining multispecies ortholog clustering and newly built, family-specific HMMs is used to identify the MAP3K gene family within seven plant species, allowing for a refinement of previously proposed gene family cladding and the novel identification of the MAP3K gene families in the allotetraploid cotton *Gossypium hirsutum* and newly sequenced monocot seagrass *Zostera marina*. The MAP3K gene family architecture is further refined and validated using bioinformatics analyses before the recently characterized Arabidopsis Raf-like MAP3K *ILK1* is identified and characterized in upland cotton. Transient gene silencing reveals an increase in RKN susceptibility following *GhILK1.1* silencing in the susceptible TM1 cultivar. No changes in susceptibility were seen in the resistant M240 cultivar or against reniform nematodes. *GhILK1.1* is only the second cotton gene characterized to have a direct role in mediating RKN resistance.

DEDICATION

I would like to dedicate this work to the family and friends who have plagued me with inexhaustible love, friendship, motivation, and happiness these last few years. Specifically, to my parents, Andrea and Tamas Bokros, who are the breathing embodiments of "you can achieve anything you truly set your mind on achieving." Without your example, sacrifice, and motivation, I would not be the person I am today – thank you. I would also like to dedicate this work to Amanda Benton, Saroj Sah, Gizem Dimlioglu, Ogunola Oluwaseun Felix, Dafne Alves, and Jeremy Winders – who are collectively responsible for years of friendship, laughter, and memories I will never forget. Thank you all for everything.

ACKNOWLEDGEMENTS

I would like to first acknowledge everyone on my committee – Dr. Sorina Popescu, Dr. George Popescu, and Dr. Martin Wubben - for the support, guidance, motivation, opportunities, and challenges you've provided these last few years; thank you for always believing in my present abilities and cultivating my future potential. I would also like to acknowledge my lab mates – Gizem Dimlioglu, Thualfeqar Al-Mohanna, Philip Berg, Setareh Nejat, and Emily Cooley - for the helpful advice, motivation, discussions, and friendships they've supported me with during my time here. Thank you to Frank Callahan for his very much appreciated help with preparing and maintaining the hundreds of plants we've worked on in this project. Thank you to Tony Arick for the help he has provided in program installations/troubleshooting and always having time to answer my impromptu questions in all matters of bioinformatics. Finally thank you to Dr. Federico Hoffman – who (despite his guaranteed criticisms against it) has been a valuable role model to observe and whose lectures, passing conversations, and shouts-down-hallways have made sure I continue striving for excellence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Since the release of the first sequenced plant genome for *Arabidopsis thaliana*, over 230 angiosperms have been completely sequenced and have their genomes deposited in publically accessible databases (1,2). Relative to Arabidopsis, most plant genomes enjoy less fidelity and are assembled as hundreds or thousands of contigs with significant segments of the genome either missing or placed in the incorrect orientation. Large genome sizes, expansive areas of highly repetitive regions, difficult to sequence heterochromatic regions such as centromeres, and polyploidy are all common bottlenecks responsible for impeding reference-quality plant genome identifications. Despite these issues, draft genomes often capture the functional gene space – or the collection of encoded genes within an organism – well and thus are excellent tools for examining an organism's phylogeny and gene family evolution (3). A comparative, multispecies examination of the evolution of functionally conserved gene families within both well-studied model plants and emerging, newly sequenced plants of agricultural and environmental importance is now feasible and would aid in the functional annotation of individual genes and complete gene families future researchers can leverage to fortify plant life against a rapidly changing global environment.

The globally important *Gossypium hirsutum* is one such commercial crop possessing a recently sequenced draft-quality genome. For decades a satisfactory

assembly of upland cotton has been difficult to resolve due to a predicted genome size of about 2.3GB (compared to 0.125GB size of Arabidopsis), an abundance of repetitive sequences (over two thirds of its genome is composed of repeated sequences), and high genetic redundancy due to an allotetraploid genome (4). Recent efforts utilizing BAC end sequencing, Illumina resequencing, physical mapping, and additional screening of small repetitive scaffolds have further improved genome annotations and resulted in the pre-release of a high-quality *Gossypium hirsutum* genome assembled into 26 synteny-oriented chromosome-level scaffolds available on Phytozome (5). Similarly, the genome of the seagrass *Zostera marina* -a uniquely evolved marine angiosperm with adaptations important for specialized ion homeostasis, nutrient uptake, and O_2/CO_2 exchange- has also been newly sequenced (6). Able to survive in a high salinity environment, comparative analyses with major crop plants and model plant species would improve our understanding of how plants have evolved to tolerate a wide range of environments.

Neither plant species presently has any known examination of its MAP3K gene family. With significant expansions across the plant kingdom, the MAP3K gene family is widely studied as its constituents and downstream interactors – MAP2Ks and subsequently MAPKs – have frequently been linked to functions in regulating development and responding to a wide range of biotic and abiotic stresses. MAPK cascades, initiated by MAP3Ks, have been shown to generate a variety of stress responses including the generation of reactive oxygen species (ROS), activating downstream defense genes (R genes), promoting the strengthening of cell walls, modulating stomatal openings to decrease the rate of pathogen entry into host plant tissues, and modifying plant growth and developments during times of stress (7,8). The

co-identification of the MAP3K gene families of *G. hirsutum* and *Z. marina* in parallel with the MAP3K gene families of relatively well-studied model organisms would allow for the novel annotation of genes involved in stress tolerance in response to otherwise damaging local stimuli.

Recently, the *ILK1* gene in Arabidopsis has been characterized as a functional Raf-like MAP3K involved in regulating homeostatic ion fluxes to respond to hyperosmotic stress and function in mediating resistance against bacterial pathogens (9). *ILK1* has been shown to function in PTI-mediated plant immunity, defined as the primary basal immune response system plants utilize to rapidly respond to environmental challenges. Further, *ILK1* operates downstream or independently of ROS production and instead is involved in ion homeostasis regulation through an interaction with the HAK5 K^+ transporter. Although *ILK1* does not appear to phosphorylate *HAK5*, *ILK1* promotes the accumulation of *HAK5* allowing cell membranes to depolarize more rapidly in response to perceived PAMPs. This depolarization is often followed by a rapid influx of extracellular Ca^{2+} ions responsible for a range of downstream signaling events (10). PTI responses work together to control early pathogenesis and the robustness of such responses ultimately decide the virulence of invading pathogens.

Although numerous genes involved in later ETI responses have been studied in response to phytonematodes, PTI responses have thus far been difficult to examine. Accurately identifying early nematode infection time points, the isolation of significant amounts of infected tissues during early stage nematode infections, and differentiating between PTI responses and later ETI responses have largely impeded the study of PTI responses against plant-parasitic nematodes (10). To the authors' best knowledge, only a

single instance of PTI responses against phytonematodes has thus far been reported. BAK1, a known interactor of FLS2, is widely known to function in flg22 (a bacterial MAMP) recognition and initiate downstream MAPK cascades ultimately promoting PTI-mediated innate immunity (11). Silencing of the *BAK1* gene in Arabidopsis has recently been shown to increase RKN susceptibility due to a decreased capacity to initiate a PTI response (12). In cotton specifically, no such interactions have been described, and only a single gene has been previously characterized as having a direct role in mediating RKN resistance. *MIC-3*, encoding a 14 kDa polypeptide lacking presently distinguishable functional motifs, was previously shown to increase RKN susceptibility in cotton 60-75% in transgenic lines (13).

In this study, the MAP3K gene families of seven plant species are characterized using a newly developed methodology combining aspects of HMMs with extensive multispecies orthogroup clustering to reveal significant expansions in agronomical and environmentally important monocots and dicots. The MAP3K gene families of *G. hirsutum* and *Z. marina* are examined for the first time, and bioinformatics analyses are used to refine previously proposed gene family claddings. Further, *ILK1* homologs in *G. hirsutum* are identified and transiently silenced to identify functional orthologs shown to function in RKN resistance. This study combines bioinformatics analyses with functional gene characterization to translate previous gene characterizations from model plant species into globally important commercial crops.

CHAPTER II

MULTISPECIES GENOME-WIDE ANALYSIS DEFINES THE MAP3K GENE FAMILY IN *GOSSYPIUM HIRSUTUM* AND REVEALS CONSERVED FAMILY EXPANSIONS

Abstract

Gene families are sets of structurally and evolutionarily related genes – in one or multiple species – that typically share a conserved biological function. As such, the identification and subsequent analyses of entire gene families are widely employed in the fields of evolutionary and functional genomics of both well established and newly sequenced plant genomes. Currently, plant gene families are typically identified using one of two major ways: 1) HMM-profile based searches using models built on *Arabidopsis thaliana* genes or 2) coding sequence homology searches using curated databases. Integrated databases containing functionally annotated genes and gene families have been developed for model organisms and several important crops; however, a comprehensive methodology for gene family annotation is currently lacking, preventing automated annotation of newly sequenced genomes. This paper proposes a combined measure of homology identification, motif conservation, phylogenomic and integrated gene expression analyses to define gene family structures in multiple plant species. The MAP3K gene families in seven plant species, including two currently unexamined

species *Gossypium hirsutum*, and *Zostera marina*, were characterized to reveal new insights into their collective function and evolution and demonstrate the effectiveness of our novel methodology. Compared with recent reports, this methodology performs significantly better for the identification and analysis of gene family members in several monocots/dicots, diploid as well as polyploid plant species.

Introduction

Mitogen-activated protein kinase (MAPK) cascades are conserved signal transduction pathways with important functions in plant growth, development, and response to environmental stresses in all eukaryotic organisms. Consequently, the identification of their members – MAP3Ks, MAP2Ks, and MAPKs - is critical to a complete understanding of how plants respond to their increasingly challenging environments. Presently, the MAP3K, MAP2K, and MAPK gene families have already been characterized in numerous plant species including Arabidopsis, strawberry, maize, canola, diploid cotton, rice, barrel clover, tomato, soybean, and grape (14–23). While MAP3Ks represent the largest, most divergent, and most poorly characterized component of the MAPK signaling cascade, continued research into how MAP3Ks function has yielded a wealth of data that has yet to be integrated into a much-needed refinement of MAP3K genomic architectures established over a decade ago (24).

As sequence data continues to accumulate for an ever increasing number of species, BLAST-based, HMM-based, and comparative homology-based searches have regularly been employed to identify entire gene families in a wide range of species. BLAST-based approaches have generally enjoyed the most popularity and involve using members of a known gene family in well-studied species to identify appropriate gene

family members in a new organism of interest based on local sequence homology. Recently, HMM-based searches have been gaining popularity and display higher accuracy in gene family identifications compared to traditional BLAST-based approaches (25). Instead of relying on individual sequences to query a database, HMM-based searches build a single probabilistic model of an entire gene family using a collection of previously validated sequences. Although both methods work well at identifying complete gene families, they also require extensive manual curation steps where hits are filtered to remove sequences that lack conserved sequence motifs or functional domains. While online databases such as Phytozome, PLAZA, and GreenPhylDB have been described as the highest performing gene family identification tool currently available (26), they often either include erroneously identified sequence hits, lack appropriate annotations necessary for accurate gene family identification, or exclude from analyses many newly sequenced species.

One of the central aims of this work is to refine the underlying architecture of the MAP3K gene family following the evolution of flowering plants. To this end, seven plant proteomes representative of the two major descendants of angiosperms, monocots, and dicots - were assembled and critically (re-) examined to identify and validate their collective MAP3K gene families. Of the five previously examined species, tomato, maize, and *Gossypium raimondii* relied on local BLAST searches using previously identified Arabidopsis, rice, and maize MAP3K sequences whereas for soybean HMM models built from a comprehensive collection of known MAP3Ks were utilized to identify new MAP3K genes. While these two methods work well at defining gene families, this study argues that a more integrative method utilizing orthogroup

identification – resulting from the OrthoMCL workflow – combined with HMMsearches can be a reliable methodology for gene family classifications. Conserved motif analysis, phylogenetic analysis, and gene duplication/collinearity analysis can further be supplemented to improve the accuracy of identifying gene family structures and provide additional functional and evolutionary insights.

This study proposes a sequential workflow where homology search is used first to decide on gene classification, followed by conserved motif analysis, phylogenetic and gene duplication analysis to define gene family evolution; a final step of integrating gene expression patterns with phylogeny offer additional functional insights and validation of gene family analysis. Figure 2.1 illustrates the workflow conducted in this study.

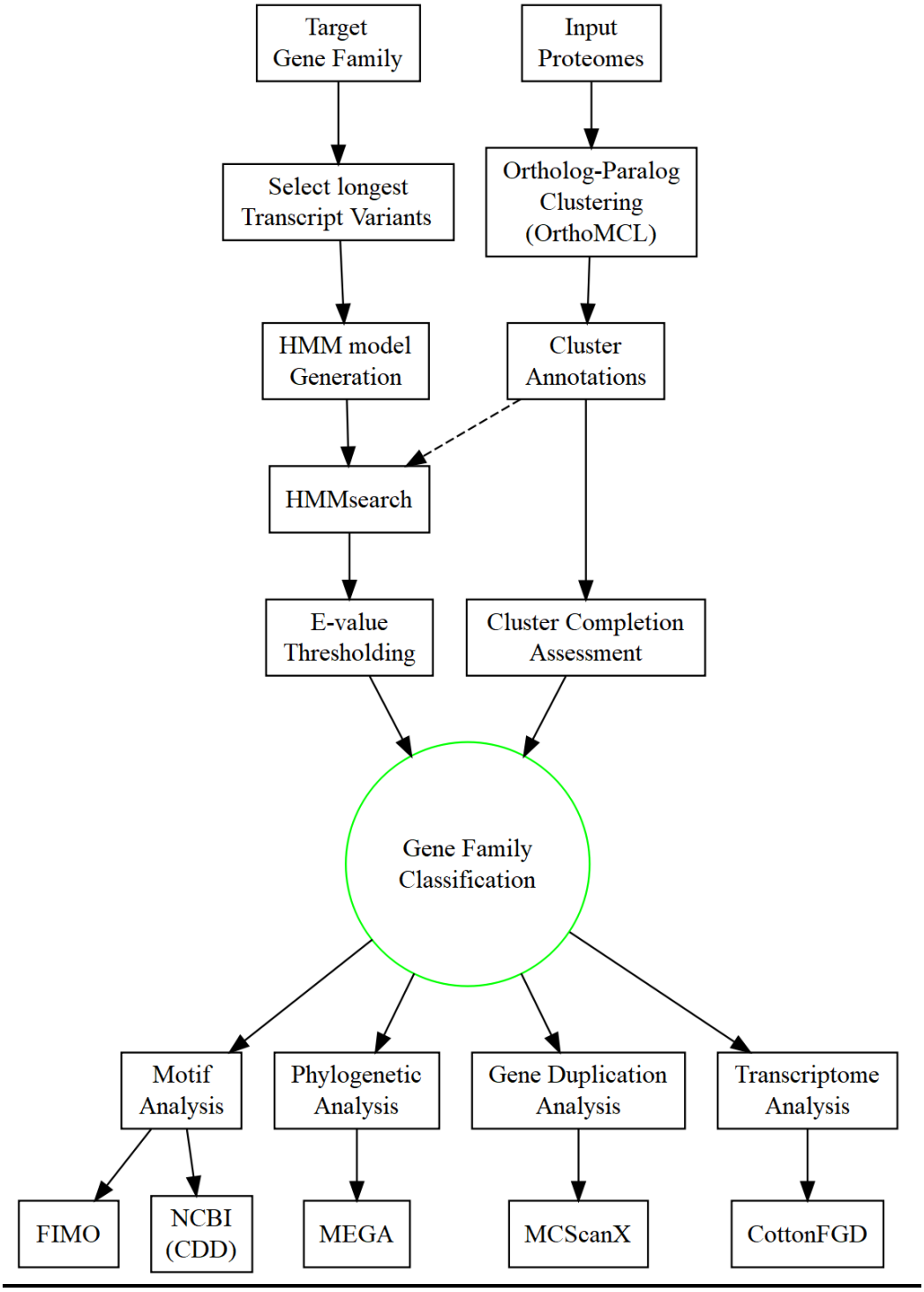


Figure 2.1 Gene family classification workflow

Flowchart illustrating steps in presented MAP3K gene family identification and subsequent analysis.

.

Results and Discussion

Cluster database construction

The proteomes of *Arabidopsis thaliana*, *Glycine max*, *Gossypium raimondii*, *Gossypium hirsutum*, *Solanum lycopersicum*, *Zea mays*, and *Zostera marina* were gathered from Phytozome and clustered into orthogroups of orthologs and recent paralogs by OrthoMCL as described below in *Methods*. 382,192 proteins from the seven proteomes were clustered into 40,524 orthogroups, excluding singletons; 63,913 unclustered proteins (singletons) were appended to this dataset to generate a final set of 104,437 orthogroups. Table 2.1 below shows the representation of each species in the 40,524 major orthogroups.

Table 2.1 OrthoMCL output clusters

Species	Input Protein Sequences	Clusters Containing Species	Unique Clusters	Absent Clusters
A. thaliana	48,456	14,687 (36.24%)	1,793	220
G. max	88,647	16,671 (41.14%)	2,603	75
G. hirsutum	87,800	23,315 (57.53%)	1,686	11
G. raimondii	77,267	23,016 (56.79%)	1,274	18
Z. mays	88,760	19,891 (49.08%)	8,115	520
S. lycopersicum	34,725	14,318 (35.33%)	1,114	244
Z. marina	20,450	10,914 (26.93%)	507	1,001

All transcript variants were retained during clustering. "Unique Clusters" depict clusters containing only sequences from a particular species, and "Absent Clusters" represent clusters missing sequences from a particular species. They are associated with orthogroups uniquely present or uniquely missing in a particular species of those examined. Singleton clusters are not included in these counts.

MAP3K identification

MAP3Ks in *Arabidopsis thaliana* (9,14), *Glycine max* (22), *Gossypium raimondii* (18), *Solanum lycopersicum* (21), and *Zea mays* (16) had previously been critically examined, while identification of MAP3Ks in *Gossypium hirsutum* and *Zostera marina* is still lacking. As described in Figure 1, our pipeline uses two homology search methods to classify MAP3K genes from the seven target species. First, we used OrthoMCL to identify orthogroups as described in the previous section. By utilizing previously identified MAP3Ks, we uncovered MAP3K orthoclusters containing sequence hits for all seven examined species. Second, we used profile-HMM homology search to identify candidate MAP3Ks from all seven target species. HMM models were built using previously identified Arabidopsis MAP3K protein coding sequences for each gene subfamily (ZIK, MEKK, RAF). The HMMsearch output and the cluster completion ratio (percentage of genes in an orthocluster above the HMMsearch threshold) were subsequently used to decide gene family membership as described in Table 2.2.

Table 2.2 Decision table used to identify gene family members

Hit is:	Above Threshold	Below Threshold
In Known Cluster	Complete clusters (Keep)	(Reject)
Outside Known Cluster	Singleton cluster (Keep)	(Reject)
	Partial cluster (Keep if cluster representation is ≥ 50%)	
	Partial cluster (Reject if cluster representation is < 50%)	

The threshold for inclusion was set as the E-value of the last identified Arabidopsis MAP3K in each HMMsearch output. A *known cluster* is any cluster that contains a previously identified MAP3K. A *singleton cluster* was defined as a cluster with only a single gene present (including clusters with multiple transcript variants of the same gene), a *complete cluster* is a cluster where all genes are above the threshold of inclusion, and a *partial cluster* was defined as a cluster with members located both above and below the threshold for inclusion. All Arabidopsis MAP3Ks from the kinome examination were kept.

In a comprehensive examination of the Arabidopsis kinome, Zulawski et al. (14) reported 48 RAFs in Arabidopsis - excluding AT2G43850 as it didn't pass their threshold for inclusion as a kinase protein – exhibiting only 69.72% similarity to a kinase HMM, below their threshold of 70%. AT2G43850, however, was included as a 49th RAF in the present study as it has recently been shown to be a functionally active RAF-like kinase involved in environmental stress responses (9).

Using this classification methodology, we categorized 108 ZIK, 255 MEKK, and 468 RAF genes. Table 2.3 gives a comparative layout of all previously and newly identified MAP3Ks; a list of all presently identified MAP3Ks with their longest transcript

variants - organized by subfamily - can be found in Appendix Table A.1. 831 MAP3Ks – containing 365 newly identified MAP3Ks - were identified within the seven examined species, including newly identified MAP3Ks within the previously unexamined *G. hirsutum* (215 MAP3Ks) and *Z. marina* (51 MAP3Ks).

Table 2.3 MAP3Ks identified in examined species

	ZIK				MEKK				RAF			
Species	**Ident.**	**Publ.**	**New**	**Total**	**Ident.**	**Publ.**	**New**	**Total**	**Ident.**	**Publ.**	**New**	**Total**
A. thaliana	11	11	0	11	37	37	0	37	49	49	0	49
G. max	24	24	4	28	34	34	23	57	84	90	6	90
G. hirsutum	N/A	N/A	27	27	N/A	N/A	52	52	N/A	N/A	136	136
G. raimondii	12	12	2	14	22	22	5	27	44	44	26	70
Z. mays	6	6	3	9	21	22	7	28	43	45	6	49
S. lycopersicum	13	16	0	13	30	33	4	34	36	40	13	49
Z. marina	N/A	N/A	6	6	N/A	N/A	20	20	N/A	N/A	25	25
Total				108				255				468

"Identified" (Ident.) column represents all genes identified that were previously published. "Published" (Publ.) column represents previously published MAP3Ks, "New" column shows MAP3Ks identified in this study but not in previous studies, and "Total" column represents all MAP3Ks presently identified.

Consistent with previous reports, the relative size of MAP3K subfamilies is similar to the one in Arabidopsis (12% ZIKs, 38% MEKKs and 51% RAFs), except for *Gossypium* where the RAFs have expanded to 63%. While ZIKs appear to have been correctly identified in previous studies, the current study proposes significant additions to both the MEKK and RAF subfamilies. Notably, the soybean MEKK subfamily has a substantial 68% increase in size compared to previous estimates while *G. raimondii* shows a large 59% increase in comparison to its reported RAF subfamily size. For the newly sequenced *Z. marina*, a total of 6 ZIKs, 20 MEKKs, and 25 RAFs were identified, while the allotetraploid *G. hirsutum* was found to encode 27 ZIKs, 52 MEKKs, and 137

RAFs - approximately double the currently identified MAP3Ks in its diploid progenitor *G. raimondii*.

We estimate that the integrative methodology has a high accuracy in assigning kinase families, as 96.1% of previously identified MAP3Ks were also currently identified. The majority of currently excluded yet previously identified MAP3Ks were located in partial clusters rejected by our decision method. These excluded sequences were classified in partially complete OrthMCL clusters with not enough agreement with the HMM search results (less than 50% of the cluster members were detected above the cut-off threshold in the HMM output). For example, some previously identified MAP3Ks in *S. lycopersicum* (21) including Solyc10g079130 (*SlMAPKKK7*6) -labeled CDPK14 in PANTHER (27)- and Solyc01g005030 (*SlMAPKKK1*) -classified as a transmembrane protein in PANTHER- showed both weak similarities to query HMM models and clustered in unselected orthoclusters. On the other hand, 90% (90/100) of the newly identified genes within examined species were found within clusters containing at least one previously identified MAP3K gene. While a more dynamic threshold of membership inclusion that takes into account motif and evolutionary analyses can be used to better resolve the most divergent members in new genomes, the presented gene family classification depicts a robust estimation of all MAP3Ks in the seven species examined.

Sequence motif analysis

Following initial gene family identification, conserved sequence motifs associated with subdomain VIII of MAP3K kinase domain were verified; this subdomain has been shown to play a major role in kinase peptide substrate recognition (28). Although variations exist in motif conservation, all 108 presently identified ZIKs were found to

have the GTPEFMAPE(L/V/M)(Y/F/L) motif conserved with a p-value < 0.0001 as calculated by FIMO. Further examination revealed that 246/255 MEKKs were found to have the G(T/S)Px(F/Y/W)MAPEV motif and 457/468 identified RAFs were found to have the GTxx(W/Y)MAPE motif similarly conserved. MAP3Ks lacking characteristic conserved sequence motifs include both previously known and newly identified MAP3Ks: 4 known and 5 new MEKKs and 5 known and 6 new RAFs. The MEKK and RAF genes not selected by FIMO analysis at the significance level of 0.0001 were found to include more divergent motifs; reducing the stringency of FIMO searches to include p-values < 0.001 resulted in only 1 known and 4 new MEKKs and 1 known RAF remaining without the presence of a detectable, diverging motif.

We also calculated the FIMO scores of the previously identified MAP3Ks rejected by our decision. We detected the conserved motifs in the three excluded ZIKs from *S. lycopersicum*, in one MEKK from *S. lycopersicum* and in 8 excluded RAFs, indicating the reason for their previous classification as MAP3Ks despite divergence from other members of their families. Altogether these indicate that another decision factor for gene family membership can be based on the results of the conserved motif search, in addition to homology searches described in our method; conserved motifs associated with kinase gene families might also, however, exhibit similarities which would reduce their discrimination power for gene family classification.

Conserved domain analysis revealed that while all ZIKs and MEKKs were found to encode only a conserved kinase domain, over half of all examined RAFs displayed secondary domains including Ankyrin repeat regions (12%), PB1 domains (13%), PAS domains (7%), ACT domains (11%) and EDR1 domains (19%). Supplementary Figure

1, 2, and 3 contain cladograms for all identified ZIKs, MEKKs, and RAFs respectively, with representations of polypeptide sequences and relevant functional motifs highlighted beside sequence identifiers. Of the 831 examined MAP3Ks only a single RAF kinase returned an unexpected secondary domain; the newly identified Gohir.D06G196600 (which also encodes a PAS domain) was predicted to encode a C-terminal truncated COG3942 superfamily domain. Interestingly, all RAFs with a divergent RAF sequence motif were found to encode secondary domains associated with suspected substrate recognition functions. Eight of the eleven MAP3Ks containing a divergent RAF sequence motif were found to encode Ankyrin Repeat domains widely known to mediate protein-protein interactions (29), while the remaining three RAFs displayed EDR1 domains which have been shown to mediate protein-protein interactions in EDR1 (30).

Phylogenetic analysis

Maximum likelihood trees were annotated with orthogroup identifiers generated by OrthoMCL. Figure 2.2 depicts a circular cladogram of all presently identified ZIKs.

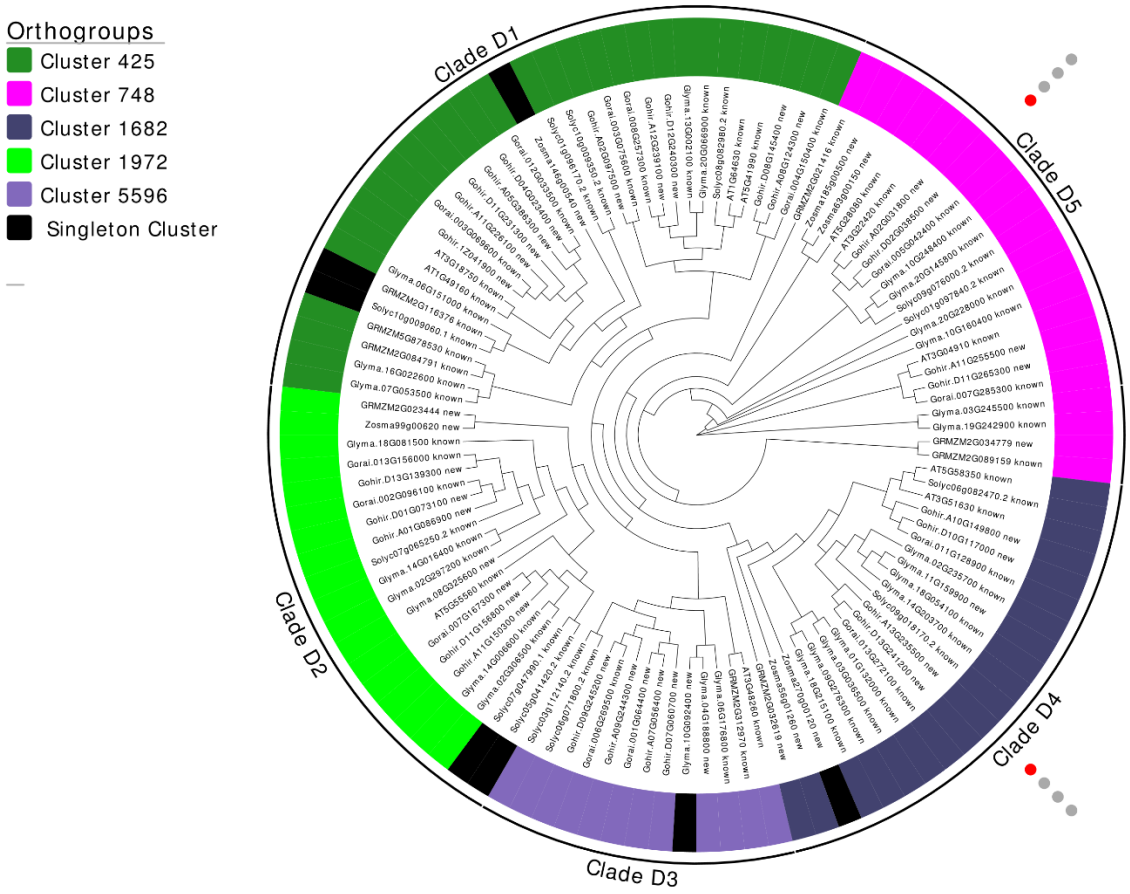


Figure 2.2 Circular cladogram of ZIK subfamily in seven plant species

OrthoMCL predicted orthogroup clusters are depicted by colored stripes beside sequence identifiers. Significantly differentially regulated genes for *Gossypium hirsutum* are depicted with circles on the outer perimeter where blue circles indicate significant downregulation, red circles significant upregulation, and grey circles are not differentially regulated. Circles represent 1) cold stress, 2) heat stress, 3) drought stress, and 4) salt stress from innermost -> outermost circle.

While previous studies (17) have predicted four major clades within the ZIK subfamily - indicative of four ancestral ZIK genes - the current study supports the existence of 5 distinct clades with five major ZIK genes present before the split of monocots and dicots. The five ZIK clades contain at least a single representative gene from at least five of the seven plant species examined – including at least one monocot and one dicot - and clustered well into the five major orthogroups depicted in Figure 2.2.

MEKKs were also previously classified into four major subclades. Orthogroup clustering and a more comprehensive phylogenetic examination support the addition of a 5th MEKK clade (Clade A5 in Figure 2.3). The new clade contains representatives from all species examined and contains about half of all presently identified MEKKs (118/255 MEKKs; 67 known and 51 new). Orthogroup cladding supports the existence of 9 ancestral MEKKs before the split of monocots and dicots, although clade A5 appears to have expanded significantly resulting in the formation of numerous paralogs within all examined species.

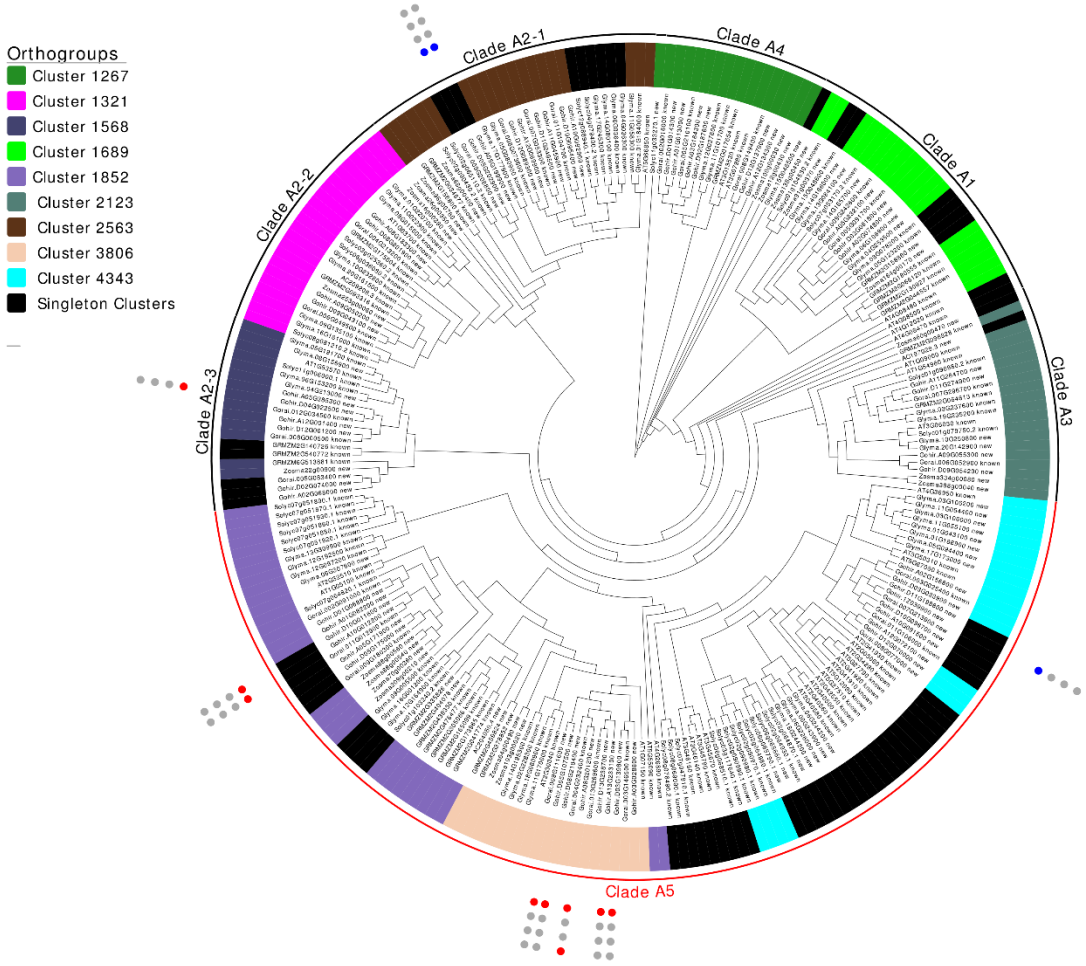


Figure 2.3 Circular cladogram of MEKK subfamily in seven plant species

Orthogroup labeling is represented by colored stripes beside leaves. Significantly differentially regulated genes for Gossypium hirsutum are depicted with circles on the outer perimeter where blue circles indicate significant downregulation, red circles significant upregulation, and grey circles are not differentially regulated. Circles represent 1) cold stress, 2) heat stress, 3) drought stress, and 4) salt stress from innermost -> outermost circle. Clade A1, A3, and A5 are expansions of their previously defined clades defined by MAPK group in 2002. Clade A2 has been split into three subclades (A2-1, A2-2, and A2-3) based on orthogroup cladding. The largely expanded clade A5 (in red) is newly proposed - residing between clades A2 and A3.

A comparison between previously suggested RAF cladding and the present orthogroup cladding reveals good conservation of major clades (Figure 2.4). However, three major refinements should be noted from the current analysis. First, while all RAF clades appear to have undergone various degrees of expansion within the examined plant species, the previously proposed clade C4 – containing ATN1-like kinases – appears to lack this expansion and is found to be constituted of only seven members, three of which are in Arabidopsis with no representation in monocots. This lack of expansion suggests a merging of clade C4 with its phylogenetic neighbor clade C3, which displays an appropriate pattern of expansion across the examined species. The two merged clades are represented by the new clade C3 in Figure 4. Second, recent domain classification in the CDD/SPARCLE database (31) indicates EDR1 functional domains in both clades B1 and B3, whereas previously EDR1 domains were only detected in clade B3. As the two clades are phylogenetic neighbors, their merger results in the combination of functionally and evolutionarily similar RAFs into a single EDR1 clade. In the present analysis, the previously proposed clades B1 and B3 are combined into the single clade B3. Lastly, AT5G07140 – classified as a MAP3K-RAF following an extensive kinome examination in *Arabidopsis thaliana* (14) – is a distinct outlier among the currently identified RAF genes. Orthogroup classification places AT5G07140 in a cluster containing AT5G58520, a predicted MAP4Ks, and no other members of this cluster were classified as RAFs in the HMM search output. Since there is no support in our analysis to include it with the RAF family, it is excluded from the currently proposed clades and instead is used to root the RAF subfamily phylogenetic tree.

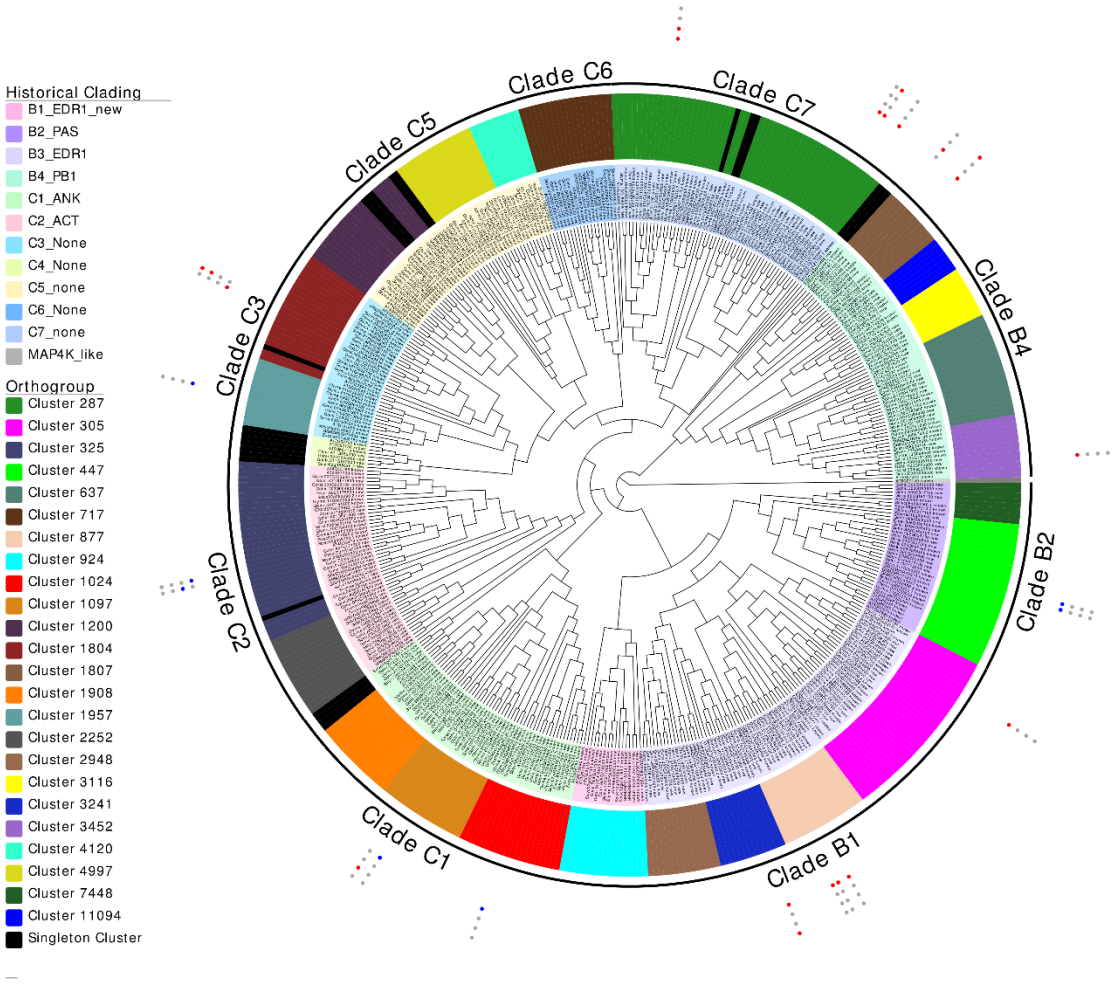


Figure 2.4 Circular cladogram of RAF subfamily in seven plant species

Differentially regulated genes for *Gossypium hirsutum* are depicted with circles on the outer perimeter where blue circles indicate significant downregulation, red circles significant upregulation, and grey circles are not differentially regulated. Circles represent 1) cold stress, 2) heat stress, 3) drought stress, and 4) salt stress from innermost -> outermost circle. Clade B3 is a combination of previously defined clades B1 and B3. Clade C3 is a combination of previously defined clades C3 and C4. Clades B2, B4, C1, C2, C5, C6, and C7 are analogous to previously established clades. Leaves are colored according to the previously proposed cladding.

Gene duplication and collinearity analysis

Gene duplication events within six of the seven species examined (*Zostera marina* was removed as its genome lacked a sufficient level of assembly) were explored by locating physical locations of MAP3K genes on individual chromosomes. Figure 2.5 depicts how gene duplication events have expanded MAP3K subfamilies within the examined plants.

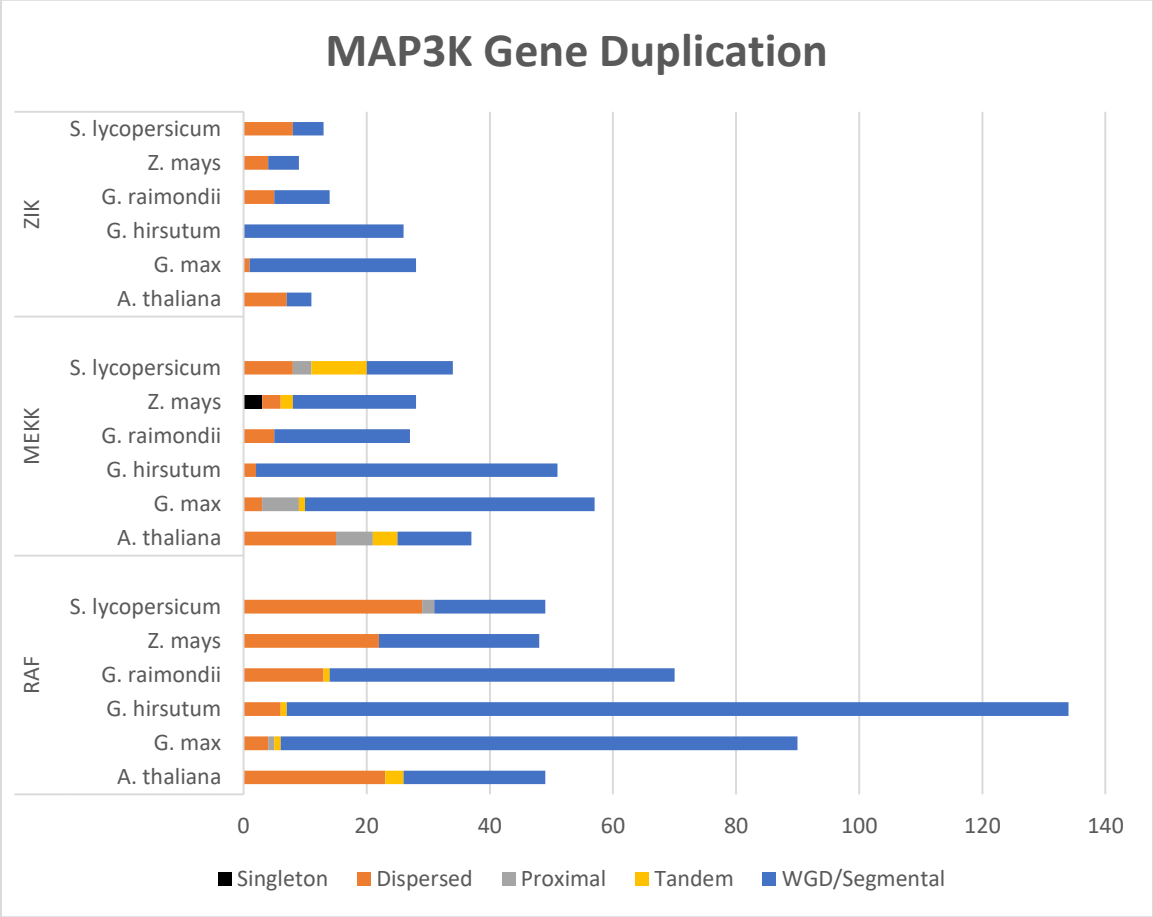


Figure 2.5 Gene duplication analysis of examined MAP3Ks

Examination of how gene duplication events have contributed to a MAP3K family expansion in 6 plants (*Zostera marina* was removed as its genome lacks sufficient assembly). WGD were responsible for the majority of predicted gene duplicates, especially in recent polyploids like *G. hirsutum* and *G. max*.

WGD/segmental (74.1%) and dispersed (20.4%) gene duplication events were primarily responsible for subfamily expansions. As expected, *G. hirsutum*, *G. max*, and *G. raimondii* consistently encoded the largest number of MAP3Ks among the species examined as they also represent the species that have undergone the most recent major gene duplication events (32).

Collinearity is a specific form of synteny requiring conserved gene order. Of the 215 MAP3Ks identified in *Gossypium hirsutum*, 211 were mapped to major chromosomes allowing for the detection of 194 collinear relationships. Collinear relationship of 133 RAFs, 43 MEKKs, and 18 ZIKs are displayed in Figure 2.6. As expected, the collinearity observed within the MAP3K family in *G. hirsutum* can primarily be attributed to its allopolyploid genome as 70.6% of collinear blocks were found between the A and D subgenome – representative of gene duplicates arising from its recent allopolyploidy. Although similar distributions of collinearity were observed within ZIKs and MEKKs, RAFs appear to have preferentially expanded within the D subgenome with almost twice (1.93x) as many collinear relationships present exclusively within the D subgenome compared to the A subgenome. These results support previous observations that the D subgenome has undergone multiple rounds of duplication and chromosomal rearrangements (33) and indicate that the large expansion of RAFs in *G. hirsutum* compared to other plants, likely resulting from its recent allopolyploidy.

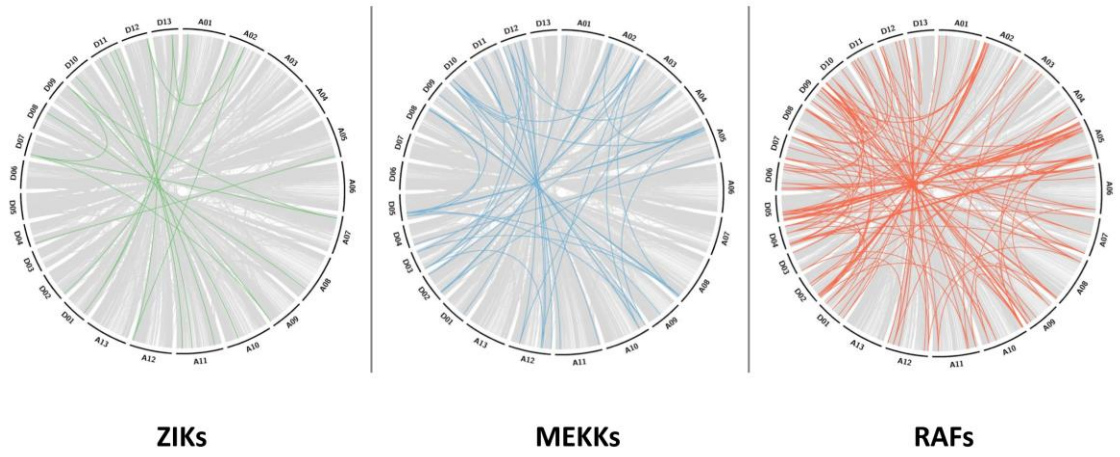


Figure 2.6 Circular collinearity plots of *Gossypium hirsutum* MAP3Ks

Subfamily-specific collinear relationships are highlighted atop a grey genomic collinearity background. While ZIKs and MEKKs display consistent collinearity between the two subgenomes, RAFs have preferentially expanded within the D subgenome.

Ka/Ks ratios were calculated to examine how MAP3Ks have diverged following duplication; *Ka/Ks* << 1 generally indicates negative or purifying selection, *Ka/Ks* = 1 indicate neutral selection and *Ka/Ks* >> 1 indicates positive selection. The *Ka/Ks* of all *G. hirsutum* MEKKs and all but one ZIK and one RAF were substantially lower than 1, with average *Ka/Ks* values of 0.31, 0.36, and 0.28 reported for ZIKs, MEKKs, and RAFs, respectively. These results indicate a strong selection bias towards gene function conservation in MAP3Ks following gene duplication events and are thought-provoking as they suggest conservation of signaling pathways in which MAP3Ks operate across the examined genomes (34,35). The observed values are consistent with previous findings that species tend to preferentially retain gene duplicates involved in signal transduction

and stress response following duplication events, increasing their environmental robustness and potential for specific adaptations (36).

Transcriptome analysis

Patterns of gene expression in *Gossypium hirsutum* were explored by querying the cotton functional genomics database CottonFGD (37) with the newly identified MAP3K genes. 2 ZIK, 11 MEKK, and 23 RAF MAP3Ks were found to exhibit at least a 50% change in expression in response to at least one biotic/abiotic stress in *Gossypium hirsutum*. These differentially expressed MAP3Ks are labeled in Figures 2.2-2.4; further examination revealed that 69.4% of identified *Gossypium hirsutum* MAP3Ks displayed at least a moderate 20% change in regulation in response to at least one of the four examined factors. Cold stress was found to induce the largest differential expression among MAP3Ks with 91.7% of all significantly differentially expressed genes showing differential expression in response to cold stress treatment. Interestingly, MEKKs were found to consistently display the largest increases in differential expression with 4 of the top 5 upregulated MAP3Ks classified as MEKKs - while RAFs consistently displayed the largest downregulation – 5 most downregulated MAP3Ks.

Of the 11 identified differentially regulated MEKKs, the newly identified clade A5 was found to contain seven upregulated, and one downregulated MEKK – strongly indicating that gene expansions within clade A5 resulted in several paralogs which might be critical participants in cold stress response. An examination of the many-to-many orthologous relationships within this clade revealed support for this claim. The closest Arabidopsis orthologs for the five upregulated *Gossypium hirsutum* genes in cluster 3806 in Figure 2.3 were found to be AT1G07150 and the *NPK1*-like AT2G30040; previous transcriptomics

analyses have revealed that both were rapidly and significantly up-regulated during cold stress responses (38). AT2G32510, the ortholog of Gohir.D10G011600 and its homeolog Gohir.A10G012200, was similarly found to be upregulated in response to cold-stress (39). Interestingly, the only MEKK downregulated in clade A5 was Gohir.1Z039900. Gohir.1Z039900 displays a many-to-many orthologous relationship with AT3G50310 (*AtMAPKKK2*0), AT4G36950 (*AtMAPKKK2*1), and AT5G67080 (*AtMAPKKK19*). While *AtMAPKKK19* was upregulated during drought and heat stress (40), functional characterization of *AtMAPKKK20* revealed cold stress-induced upregulation (41) indicative of neofunctionalization resulting from diverging homology within some of the paralogs in Clade A5.

C-clade RAFs appear to be involved in osmotic stress sensitivity. *G. hirsutum* RAFs within Clades C1, C2, C3, and C7 displayed significantly differential responses to osmotic stress conditions compared to non-stressed controls (Figure 2.3). Although not significantly upregulated, Gohir.A10G009900 located in Clade C5 was previously identified as a *MAPKKK* involved in drought, salt, and cold stress response (42) and found to be moderately upregulated (20.8%) in response to cold stress and downregulated in response to drought stress (22.5%) in the present transcriptome analysis. Clade C6, likewise, contains three *G. hirsutum* genes displaying moderate differential regulation in response to the examined osmotic stresses. Previous functional characterizations of At1g62400 (*HT1*; Clade C5), At4g18950 (*BHP*; Clade C1), At2g17700, At4g35708, At4g38470 (*STY8*, *STY17*, and *STY46* respectively; Clade C2) and At2g43850 (*ILK1*; Clade C1) further support the role in osmotic stress control for C-clade RAFs (9,43–45). Gohir.A08G065000 in Clade C7 was functionally characterized as *GhMAP3K6*5 and

found to be involved in pathogen and heat stress susceptibility (46). *GhMAP3K65* was also identified as highly upregulated during cold stress treatment. Gohir.D12G274200 in Clade C7 was recently functionally characterized in transgenic *Nicotiana benthamiana*; transgenic plants overexpressing *Gohir.D12G274200* showed increased pathogen susceptibility and improved tolerance to drought and salt stress at the seedling stage (47).

Conclusion

Our gene family definition method integrating orthologs clustering and profile HMM homology search was in very good agreement with previous large-scale studies on defining gene families in plants. However, significant differences were detected when compared with studies focusing on MAP3K genes in recently sequenced organisms (*S. lycopersicum, G. raimondii*, and *G. max*). That may be due to inherent difficulties in using a single homology search method and in defining adequate threshold levels for gene family definition. Here is where the integration of the whole genome with profile-based homology search methods provided an adequate set of rules for gene family definition. Also, conserved motif analysis, phylogenetic analysis, and gene duplication/collinearity analyses allowed for a better definition of gene clades, using evidence from gene family evolution and functional motif conservation. Large changes from the previously reported MAP3K families were found in the expanded subfamily of RAFs (*G. raimondii* and *S. lycopersicum*) but also in the more conserved MEKK subfamily (*G. max*).

Compared to previous estimates, within the re-examined species, newly identified MAP3Ks account for a significant change in MAP3K family size. While previously having gone unreported, the newly identified MAP3Ks consistently display phylogenetic similarity and high sequence homology to known MAP3Ks and encode subfamily specific

motifs and functional domains, indicating potential shared functional equivalency. Although current findings provide an accurate assessment of MAP3Ks in seven plant species, improvements in gene family member identification could be achieved with the application of a more dynamically inferred threshold; potentially one defined using a machine learning algorithm.

Significant expansions within the MAP3K gene family have been uncovered following an extensive examination of plant monocots and dicots. These findings allowed for refinement of the previously proposed MAP3K family cladding. A more comprehensive sampling of plant species, extensive ortholog clustering, functional domain characterization, and subfamily specific, HMM-based homology assessments allowed for a robust definition of MAP3Ks.

In the diverging MEKK and RAF subclades, conserved subclade functionality is supported by transcriptomic evidence, recent gene characterization studies, and orthogroup clustering. Clade C RAFs were consistently found to be differentially regulated in response to osmotic stresses – indicating their likely roles in osmotic stress responses; these roles were found to be supported by gene characterization studies of individual MAP3Ks in both *G. hirsutum* and *A. thaliana*. The newly identified Clade A5 also displayed a conserved role in cold stress response, with support from studies in Arabidopsis. A more extensive examination of MAP3Ks is needed to associate functionality with subfamily cladding. Identification of differentially regulated MAP3Ks can be used to detect targets of significant importance in plant stress response.

Gene duplication and collinearity analyses showed that MAP3Ks had expanded primarily due to WGD events. In *G. hirsutum*, collinearity analysis revealed that while

good collinearity was maintained between the A and D subgenomes in MEKKs and ZIKs, gene duplications within its D subgenome had an increased contribution to the expansion of RAF subfamily.

The work presented in this study provides an extensive examination of how MAP3Ks have expanded in plants and for the first time establishes the MAP3K gene family in the commercially important *G. hirsutum* as well as the recently sequenced monocot *Z. marina*.

Methods

Sequence retrieval, database construction, and MAP3K identification

To perform multispecies MAP3K analyses, the complete proteomes of *Arabidopsis thaliana* (48), *Gossypium raimondii* (49), *Gossypium hirsutum* (4), *Solanum lycopersicum* (50), *Glycine max* (51), *Zostera marina* (6), and *Zea mays* (52) were retrieved from Phytozome v12 (5). The sequences were uploaded onto the Cyverse Discovery environment and clustered into orthogroups using the OrthoMCL workflow detailed at (https://pods.iplantcollaborative.org/wiki/pages/viewpage.action?pageId=12881253). Using the default E-value cutoff, the top 300 hits and alignments of each query were retained as input into the OrthoMCL pipeline (53). The "OrthoMCL v1.4" application was used to cluster orthologs; index mode was set to all, a p-value cutoff of 1.5, percent identity cutoff of 0, percent match cutoff of 0, a maximum weight of 350, and an inflation parameter of 1.5 were used for clustering. Orthogroups were generated by querying the output file using "queryOrthoMCL".

Subfamily-specific HMMs and HMMsearches were built and run using HMMER 3.1b2 available at http://hmmer.org/. The threshold used for each MAP3K subfamily is

defined as the first instance of a transcript variant of the lowest scoring member of a particular subfamily in Arabidopsis. For the present analysis, AT5G28080.1, AT2G40500.1, and AT5G07140.1 were used for ZIKs, MEKKs, and RAFs with E-values of 3.40E-107, 4.3E-66, and 4.10E-79 respectively. Previously published subfamily members from *Arabidopsis thaliana* (14), *Glycine max* (22), *Gossypium raimondii* (18), *Solanum lycopersicum* (21),and *Zea mays* (16) were used in order to compare how well the decision tree performed in selecting appropriate subfamily members and are represented in "Published" columns of Table 2.3.

Sequence motif analysis

Conservation of subfamily specific sequence motifs was performed using FIMO, part of the MEME suite of tools; unique hits in individual proteins with a p-value < 0.0001 were associated with motif conservation (54). Default parameters were used to query NCBI's Conserved Domain Database (CDD) search tool's v3.16 database to identify conserved domain motifs (31).

Phylogenetic analysis

Multiple sequence alignments of subfamily specific MAP3Ks were performed using MUSCLE (55). Maximum likelihood trees for all alignments were built using MEGA v7.0.26 (56). Best fit models were predicted using the model prediction tool in MEGA, and maximum likelihood trees were built with default parameters and supported by 200 bootstrap replicates. Tree visualization was generated using Evolview (57). Cladding for all trees was based on orthogroup clustering generated from OrthoMCL and by using suggested cladding from the MAPK group (MEKKs and RAFs) (24); all depicted

major orthoclusters contained at least 5 of the 7 examined species. All trees are available online at http://120.202.110.254:8280/evolview/#shared/SibKukloHk/723.

Gene duplication and collinearity analysis

Gene duplication and collinearity analyses were performed using MCScanX using GFF3 files retrieved from Phytozome (58). Genes that lacked placement on major chromosomes were excluded from examination. Collinearity circle plots were generated using Circos v0.69 (59).

Transcriptome analysis

FPKM normalized gene expression data for *Gossypium hirsutum* (NAU) was downloaded from CottonFGD (37). Sequences had to be migrated between JGI and NAU datasets using BLAST. Query sequences from JGI were BLASTed against NAU sequences with unique best hit matches kept for further analysis; if a query had a duplicate hit, the hit on the same subgenome chromosome was used if no conclusive hit was found the gene was removed from the further analysis). Gene expression data for 22 ZIK, 47 MEKK, and 124 RAFs were identified in *Gossypium hirsutum*. For each gene, the log2(FPKM+1) of each stress treatment (cold stress, heat stress, drought stress, and salt stress) was subtracted from the log2(FPKM+1) of the control treatment to calculate the log2(Fold Change) in each treatment. Genes that displayed 0.5 > Fold change > 1.5 were labeled as significantly differentially expressed in appropriate cladograms.

CHAPTER III

IDENTIFICATION OF THE FIRST *GOSSYPIUM HIRSUTUM* MAP3K INVOLVED IN ROOT-KNOT NEMATODE RESISTANCE

Abstract

Functional gene characterization of the *ILK1* gene - recently characterized in Arabidopsis as a source of plant osmotic and pathogen stress resistance- was undertaken in *Gossypium hirsutum*. The previously developed cotton leaf crumple virus (ClCrV) viral induced gene silencing (VIGS) vector was demonstrated to effectively transiently silence target gene expression in cotton roots and used to characterize a functional ortholog of the recently characterized Arabidopsis *ILK1* gene. *ILK1* homologs were first phylogenetically identified and then transiently silenced in *Gossypium hirsutum*. Silencing of the *GhILK1.1* set of homeologs was found to significantly increase root-knot nematode (RKN, *Meloidogyne incognita*) susceptibility in the susceptible TM1 cultivar 2-8 fold. Silencing of the same set of homeologs did not cause a significant change in susceptibility to reniform nematode (*Rotylenchulus reniformis*) susceptibility in TM1 or the previously resistant M240 cultivar. Our results support the identification of the first MAP3K gene involved in basal RKN resistance in cotton. Future work identifying an associated signaling pathway could uncover new sources of improved innate host plant resistance against a broad spectrum of plant pathogens.

Introduction

With susceptible host plants ranging over 5500 different species – including all 25 of the most produced commercial crops – the root-knot nematode (RKN; *Meloidogyne incognita*) has been labeled the most widespread and damaging obligate plant parasite responsible for an estimated $100 billion loss/year worldwide (60). Although previous yield losses attributed to RKN in untreated, susceptible cotton crops have been estimated at 26%, in Mississippi only an estimated 6.1% loss in yield for the 2014 growing season attributed to RKN damage (61,62). Carefully planned management strategies involving appropriate crop rotation, nematicides, and the planting of resistant lines are common strategies for dealing with harmful nematode populations. Often, however, the difficulty of designing practical, efficient, and profitable crop rotations, the phasing out of environmentally damaging nematicides from the market, and limited availability of broad-spectrum phytonematode resistant lines with acceptable yield potential and fiber quality hinder the global potential of currently available nematode control strategies (63). Furthermore, despite available management options, yield losses in cotton due solely to nematode damage within the American cotton belt, have steadily been trending upward from 1% in 1987 to 5.6% in 2006 (http://www.cotton.org/tech/pest/nematode/losses.cfm). An understanding of early-stage pathogen resistance pathways would aid in the integration of improved, innate host plant resistance to phytonematodes and is the safest, most robust, and most economically attractive option for the future of phytonematode management, both locally and globally where local yield losses in cotton of 18-32% have previously been attributed to nematode damage (64).

Root-knot nematodes navigate the intercellular regions of host plant cells during the early J2 stage of development (10). Female root-knot nematodes reproduce asexually and establish a permanent feeding site within a cluster of cells near the root vascular cylinder. Using a stylet, the nematode injects secretory proteins into neighboring cells stimulating a change in mitotic growth regulation causing surrounding cells to continue growing uninhibited into structures known as "giant cells." These giant cells act as nutrient sinks, providing the nematode with a supply of food needed to develop eggs that eventually hatch and restart the parasitic life cycle. This process saps nutrients from the host plant, results in the formation of root galls, and significantly reduces crop yield outputs (65).

Before giant cell formation, RKN utilizes a predicted arsenal of 61 cell wall-degrading carbohydrate-active enzymes (CAZymes) to invade plant root tissue (66). Although this tissue damage is likely to produce damage-associated molecular patterns (DAMPs) and induce an early stage PTI-like basal defense response, multiple difficulties presently impede the necessary examinations required to characterize these interactions between host plants and RKN (10). As a result, characterized PTI responses have yet to be identified against phytonematodes in most plant hosts (67). Studies in other pathosystems, however, predict a likely involvement of PTI responses.

The only known instance of a characterized phytonematode-induced PTI response was documented in Arabidopsis against RKN involving the canonical PTI signaling interactor BAK1, a known FLS2 interactor involved in flg22 induced PTI responses (12). In the study, BAK1 was shown to participate in canonical PTI pathways, phosphorylating BIK1, which in turn phosphorylates a respiratory burst NADPH oxidase D (*RBOHD*) that

enhances ROS generation. Interestingly, RKN perception was also shown to not involve the MAMP recognition receptor FLS2, suggesting an immune response pathway distinct from the recognition of bacteria on the surface of nematodes. Further, of the three DAMP receptors examined – PEPR1, PEPR2, and DORN1 – none had a significant effect on altering nematode susceptibility (12). Despite this, nematode pathogenesis surely produces a range of DAMPs as invading nematodes force their way into root vascular tissues and further into root cell cytosols.

Following the perception of DAMPs, an assortment of downstream signaling events – including the initiation of mitogen-activated protein kinase (MAPK) cascades - work in unison to generate a basal plant immune response. Canonical MAPK cascades are composed of sequential activation of MAP3K-MAP2K-MAPK kinases and work to transfer and amplify signals through a cell resulting in transcriptional reprogramming able to respond to a wide array of stresses (68). Conserved plant MAPK cascades have been implicated in numerous immune responses to biotic and abiotic stresses through the induction of oxidative bursts, ethylene production, the expression of defense-related genes, and cell wall modifications (7).

In plants, MAP3Ks have expanded significantly compared to their metazoan counterparts and are the primary components of evolutionarily conserved MAPK cascades. A number of MAPK cascades have previously been identified in plants, the most well studied of which include the MAP3K *EDR1* which has been shown to interact in a cascade with *MKK4*/*MKK5*-*MPK3*/*MPK6* and negatively regulate plant defense responses and cell death (30). More recently, the Raf-like Integrin-Linked Kinase 1 (*ILK1*) has been identified as a functional MAP3K protein kinase involved in

hyperosmotic stress sensitivity and resistance against bacterial pathogens in Arabidopsis (9). In Arabidopsis, *ILK*s have expanded into a small subfamily of Raf-like protein kinases composed of 6 distinct homologs (*ILK*s1-6), all with unique transcriptional expression patterns and localizations, but with distinctly conserved structural similarities including a conserved kinase domain and up to three sets of Ankyrin repeat regions (69). These domains within the Arabidopsis *ILK1* have been linked with functions in plant cell stress responses and support their roles in signaling pathways regulating cellular homeostasis and pathogen immunity (9).

Two distinct characteristics of the recently characterized Arabidopsis *ILK1* gene position it as a promising candidate for regulating PTI responses against phytonematodes in cotton. First, *ILK1* in Arabidopsis has been shown to function in PAMP-triggered plant immunity (PTI). Through interactions with the calcium sensor CML9 and K^+ transporter *HAK5*, *ILK1* has been shown to alter plasma membrane polarity following recognition of MAMPs to regulate downstream MAPK cascades, positively regulating plant innate immunity. Similar signaling cascades are likely to exist in response to phytonematode perception in cotton, with an equivalent cotton ortholog of *ILK1* functioning in defense against nematode infection. Second, Arabidopsis with defective *ILK1* transcription showed significantly altered accumulation of multiple nutrients within cells including potassium, manganese, magnesium, sulfur, and calcium following PAMP perception. Cotton *ILK*s may also function in pathways controlling the flow of nutrients within the plant. Disruption of nutrient flow during giant cell formation may impair early nematode feeding site establishment, altering nematode susceptibility.

In this study, the *Gossypium hirsutum* ortholog of the Arabidopsis *ILK1* is first identified by homology to the recently characterized Arabidopsis *ILK1* and subsequently through functional characterization in cotton seedlings. Transient silencing of an orthologous set of upland cotton *ILK1* homeologs was shown to increase cotton susceptibility to RKN 2-8 fold compared to empty vector control plants; however, silencing of the same set of homeologs did not affect susceptibility in the RKN resistance line M240, nor did silencing alter susceptibility to reniform nematodes. From these data, we conclude that *ILK1* orthologs in upland cotton likely play a central role in basal plant resistance specifically against RKN.

Results

Assessment of inoculation methods and generation of *GhAct7* control

Viral-induced gene silencing (VIGS) was performed using the previously developed cotton leaf crumple virus (ClCrV) binary vector carried within *Agrobacterium* strain Gv3101. A construct silencing the previously characterized cotton *GhAct7* gene was generated(70). VIGS construct generation is detailed below in Methods. Two different *Agrobacterium* mediated silencing protocols -agro infiltration of cotyledons and agro-drenching of the plant soil substrate - were also tested to examine which produced the strongest silencing phenotype (Figure 3.1).

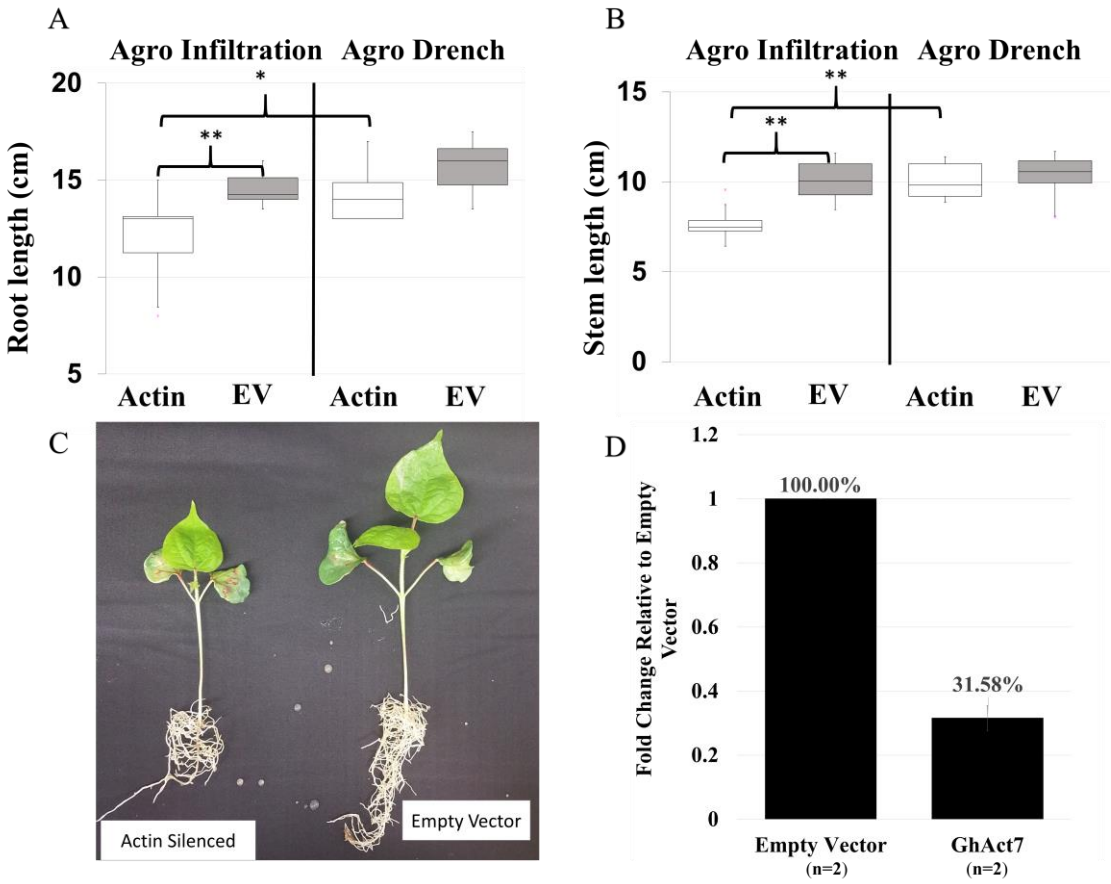


Figure 3.1 *GhAct7* silencing in upland cotton

In the box-and-whisker plots, all values are representative of measurements from 8 individual cotton seedlings 13 days after silencing. * indicates significance at a p-value of at least 0.05. ** indicates significance at a p-value of at least 0.01. "Actin" indicates *GhAct7* silenced plants and "EV" indicates empty vector control plants. A red star indicates an outlier. A) Comparison of root length, B) Comparison of stem length, C) Visual phenotype of *GhAct7* silencing in stem and roots of silenced and control cotton plants. D) Change in expression of *GhAct7* gene in silenced and empty vector control plant.

The agro-infiltration method performed significantly better than the agro-drench method following comparisons of seedling stem and root lengths between empty vector control (EV) and *GhAct7* silenced (Actin) treatments (Figure 3.1a and Figure 3.1b). Although no significant difference was observed among seedlings drenched with VIGS constructs targeting the cotton actin gene or seedlings infiltrated/drenched with empty

vector constructs (ANOVA calculated p-value of 0.1), actin infiltrated seedling root lengths were significantly shorter than empty vector infiltrated root lengths with an average length reduction of 2.5cm (p value 0.01). Similarly, stem heights were significantly shorter in actin infiltrated plants compared to empty vector infiltrated plants with a height reduction of 2.4cm (p-value 0.0005). Drenching did not result in significantly reduced root length or stem height (p-value > 0.05) but did result in an average root length reduction of 1.4cm and an average stem height reduction of 0.3cm. Actin infiltrated samples were checked for transient gene silencing using qRT-PCR and were found to display an almost 70% reduction in *GhAct7* gene expression compared to empty vector inoculated seedlings (Figure 3.1d). For all successive experiments, the agro-infiltration method was utilized.

Identification of cotton ILKs

Previously, sequence homology, ortholog grouping, functional motif analysis, and phylogenetic analyses have placed *ILK1* as a plant Raf-like MAP3K. In order to investigate the role of *ILK1* as an equivalent regulator of plant stress responses in *G. hirsutum* as in Arabidopsis, the *ILK* subfamily of clade C1 MAP3Ks was identified in Arabidopsis, *G. hirsutum*, and *G. raimondii* (Figure 3.2). Although previously six *ILK*s were identified in Arabidopsis, the *ILK* subfamily has expanded to include 20 distinct *ILK* genes in *G. hirsutum* (10 sets of homeologs) as well as 10 *ILK* genes in *G. raimondii*, which invariably aligned with a D subgenome homeolog within *G. hirsutum*.

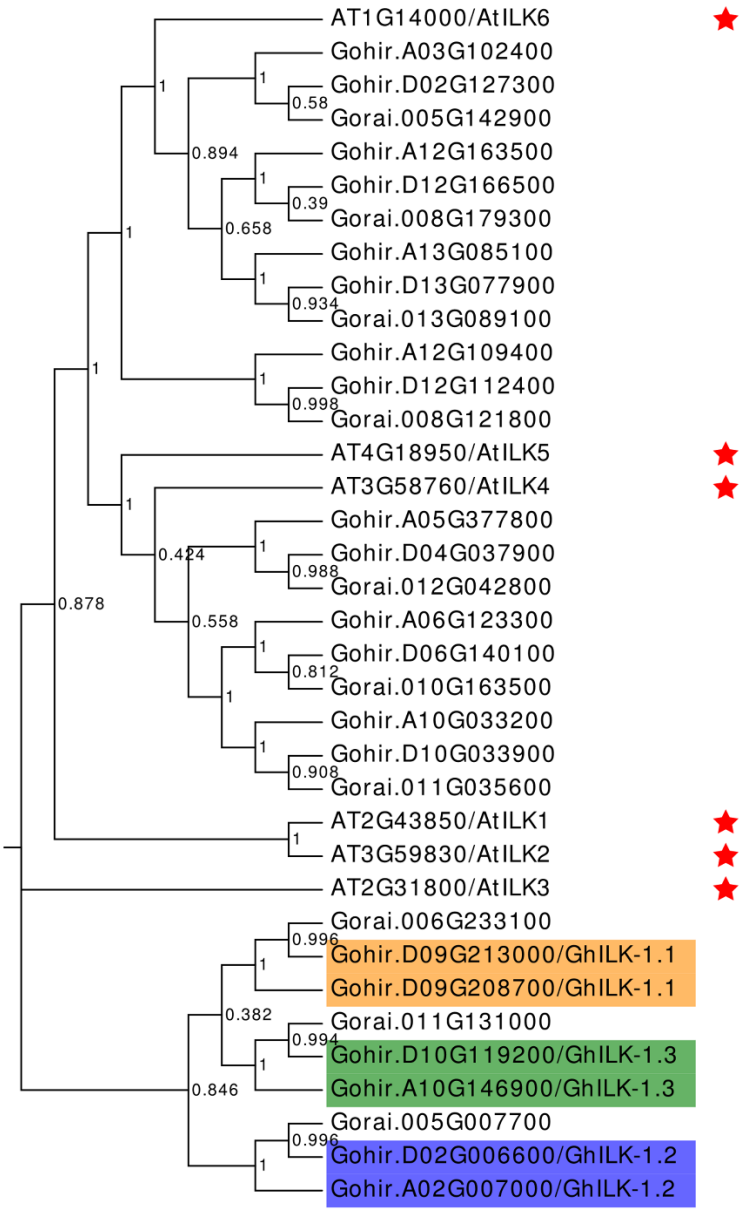


Figure 3.2 Maximum-likelihood tree of *ILK* subfamily in Arabidopsis, *G. raimondii*, and *G. hirsutum*

Presently examined sets of *ILK1* homeologs in *G. hirsutum* are colored identically; a red star is beside all previously identified Arabidopsis *ILK*s. Sequences were identified from preceding MAP3K examination. 500 bootstraps were performed to test phylogeny; bootstrap values are shown at all nodes.

OrthoMCL was used to cluster all proteins within seven plant species into clusters of orthologs and recent paralogs, classifying three of the six Arabidopsis *ILK*s (*ILK1*, *ILK2*, and *ILK3*) into a single orthocluster. All members of this *ILK1* orthocluster shared a high degree of sequence homology, although the Arabidopsis *ILK3* gene was a distinct outlier among examined sequences despite conservation of both identified Ankyrin repeat regions and a kinase domain shared among the Arabidopsis *ILK*s (Figure 3.3). Interestingly, while both monocots examined - *Zostera marina* (seagrass) and *Zea mays* (maize) – have *ILK1*-like homologs, the identified homologs have fewer paralogs compared to dicots and less conservation of the Ankyrin repeat region preceding the kinase domain.

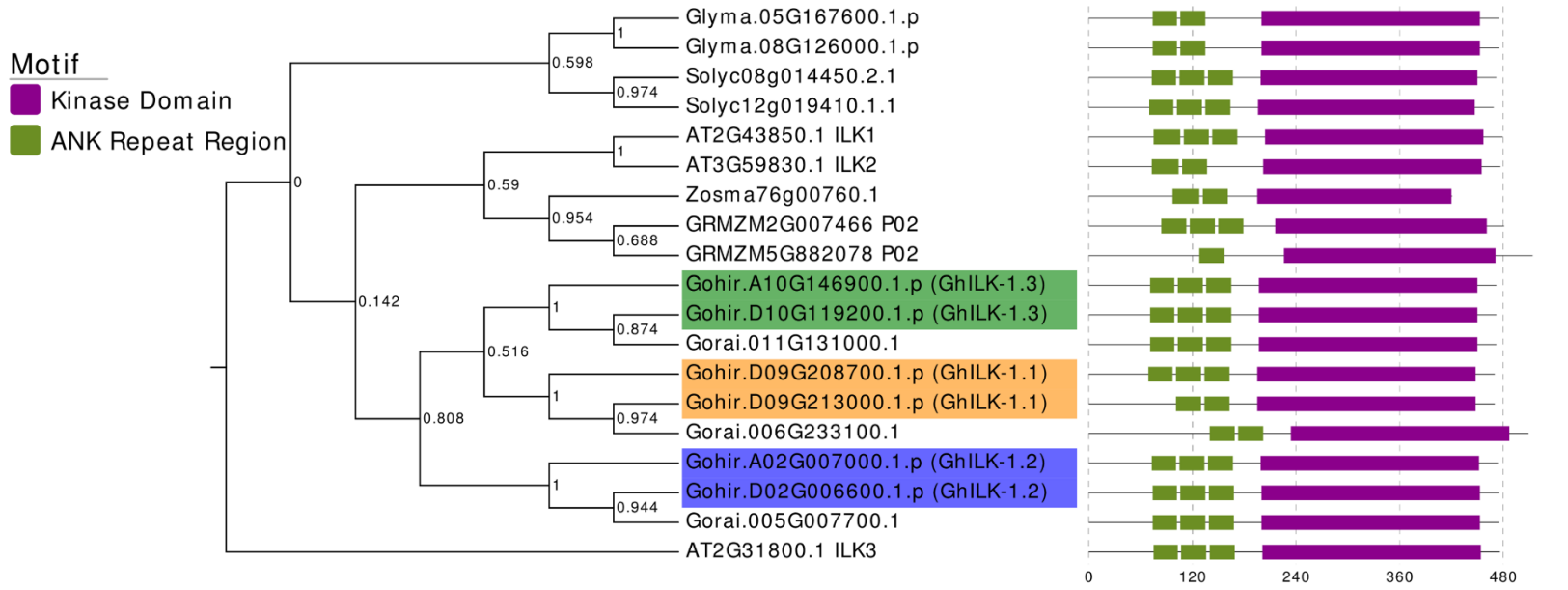


Figure 3.3 Maximum likelihood tree of *ILK1* orthocluster for seven plant species.

Sets of predicted homeologs of *ILK1* in *G. hirsutum* are highlighted and labeled with depictions of SMART-predicted domain motifs beside leaf labels. Bootstrap values from 500 bootstrap replicates are depicted at all nodes.

An examination of subdomains within the kinase region of Arabidopsis and *G. hirsutum ILK1* homologs revealed species-specific modifications to sequence motifs typically conserved among eukaryotes (Table 3.2).

Table 3.2 Conservation of Catalytically important residues in Arabidopsis and *G. hirsutum ILK*s

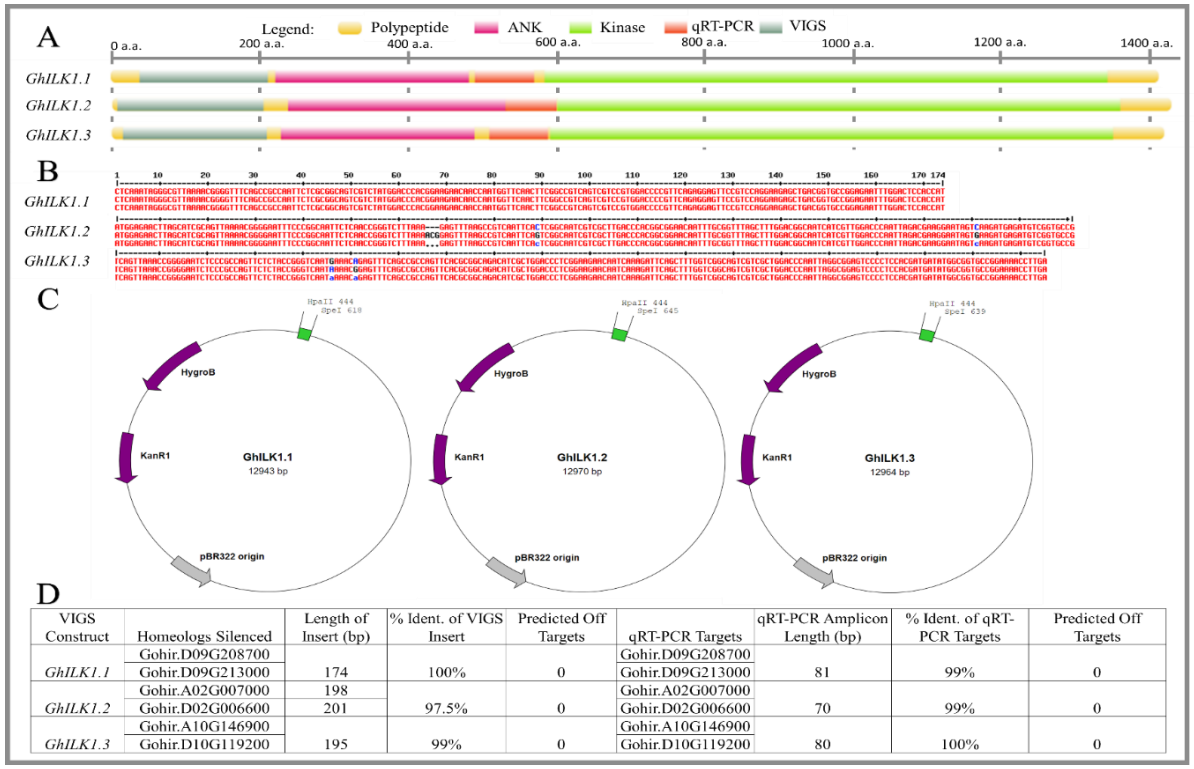
Kinase Sub-Domains	G-Loop I	Cat. Lys.	C-Loop	A-Loop
EK Consensus	G-G--G	K	HRDL---N	DFG
ILK1	S-G--Q	K	HCDL---N	GFG
ILK2	S-G--Q	K	HCEL---N	GFG
ILK3	S-G--Q	K	HCDL---N	GFG
GhILK1.1	T-G--Q	K	HCDL---N	GFG
GhILK1.2	T-G--Q	K	HCNL---N	GFG
GhILK1.3	T-G--Q	T	HCNL---N	GFG

Differences in subdomain conservation between Arabidopsis ILKs1-3 and *G. hirsutum* candidate *ILK1*s. The eukaryotic consensus is shown for comparison at the top. Residues known to be important for catalytic activity are shaded in red.

One interesting observation in Table 3.2 revolves around the conservation of catalytically important residues in specific subdomains of the examined cotton *ILK*s. *GhILK1.1* appeared to retain conservation of all three catalytically important residues within three examined kinase subdomains, however, *GhILK1.2* only retained two of the three residues while *GhILK1.3* retained only a single conserved residue. Further, an examination of previously described substrate recognition motif among Arabidopsis and *G. hirsutum ILK*s revealed modifications from the conserved GTxx(W/Y)MAPE motif. The canonically conserved APE motif, however, remained in all examined homologs.

Generation of *GhILK* VIGS silencing constructs

Although *GhILK1.1* appears to be the closest ortholog of the Arabidopsis *ILK1* gene based on conservation of catalytically important amino acid residues, all *G. hirsutum ILK* genes in the *ILK1* orthocluster were selected for functional characterization. Fragments suitable for VIGS silencing were identified as described below in Methods and amplified from root DNA using primers found in Tables B.2-4; target fragments were checked using gel electrophoresis to ensure the correct sized fragment was being amplified. Fragments unique to each set of homeologs were ligated into a modified cotton leaf crumple virus VIGS vector (Figure 3.4)(71). Insert fragments were checked for specificity to each set of target homeologs (Figure 3.3b) and checked for potential off targets. No potential off-targets were identified resulting from any insert fragment against the AD1-NBI v1.1 annotation of the *Gossypium hirsutum* genome with a mismatch allowance of 0 along a 21bp n-mer.



VIGS Construct	Homeologs Silenced	Length of Insert (bp)	% Ident. of VIGS Insert	Predicted Off Targets	qRT-PCR Targets	qRT-PCR Amplicon Length (bp)	% Ident. of qRT-PCR Targets	Predicted Off Targets
	Gohir.D09G208700				Gohir.D09G208700			
GhILK1.1	Gohir.D09G213000	174	100%	0	Gohir.D09G213000	81	99%	0
	Gohir.A02G007000	198			Gohir.A02G007000			
GhILK1.2	Gohir.D02G006600	201	97.5%	0	Gohir.D02G006600	70	99%	0
	Gohir.A10G146900				Gohir.A10G146900			
GhILK1.3	Gohir.D10G119200	195	99%	0	Gohir.D10G119200	80	100%	0

Figure 3.4 Sequence analysis of *GhILK* silencing constructs

A) Diagram of polypeptide sequences used for VIGS silencing of *ILK1* homologs in upland cotton. B) Multiple sequence alignment of VIGS insert fragment for all *ILK1* homoeologs in *G. hirsutum*. Identical residues are shown as uppercase red nucleotides and mismatch residues are shown as lowercase blue nucleotides. C) Plasmid maps generated using http://rf-cloning.org/savvy.php. VIGS insert is shown in green ligated into pJRT.Agro.ClCrVa.008 plasmid to generate silencing constructs. D) Examination of VIGS silencing regions. No off-targets were identified with zero mismatches allowed against the NBI-v1.1 *Gossypium hirsutum* database. qRT-PCR primers were BLASTed against JGI *G. hirsutum* database to identify amplicons and off-targets. All primers for both VIGS and qRT-PCR are available in Tables B2-4.

Characterization of the role of *ILKs* in plant resistance to nematodes

The role of *G. hirsutum ILK*s was examined in the susceptible cotton cultivar TM-1 and the RKN-resistant line M-240. In an initial examination of the functions of *ILK1* candidate orthologs, susceptibility was found to be increased in *GhILK1.1* silenced TM1 plants. *GhILK1.1* silenced seedlings were the only plants to show a significant change in RKN susceptibility following targeted gene silencing - displaying an 8-fold increase in susceptibility compared to empty vector controls displaying uninhibited target gene expression. Silencing of *ILK1* homologs varied in each construct, with *GhILK1.1* displaying the strongest silencing at around a 90% decrease in expression compared to empty vector controls (Figure 3.5c). Transient gene silencing of *GhILK1.3* was more moderate at around 66%, and *GhILK1.2* plants did not show any gene silencing compared to empty vector inoculated controls. Silencing of all homeologs in a single plant (*ILK* Mix in Figure 3.5a) did not induce transient gene silencing able to significantly alter RKN susceptibility (Figure 3.5b). Changes in susceptibility were also assessed in the nematode resistant M240 line with no significant change in susceptibility observed (Figure not shown). As many constructs were tested and sample sizes were small for both gene silencing assessments and RKN susceptibility assays, the experiment was repeated focusing exclusively on *GhILK1.1*. Future experiments confirming efficient silencing of *GhILK1.2* and *GhILK1.3* in more plants subjected to RKN are necessary to accurately assess the potential functionality of these sets of cotton *ILK1* paralogs.

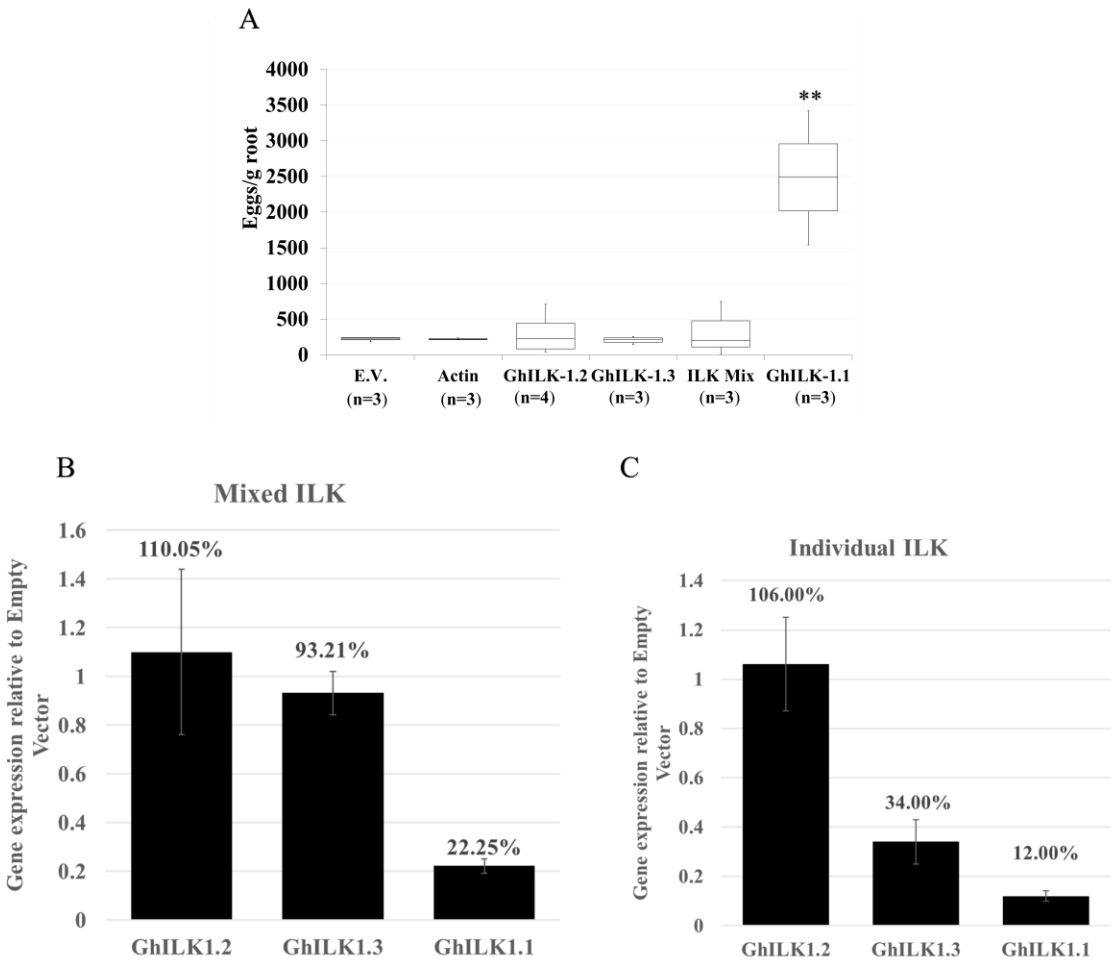


Figure 3.5 RKN susceptibility of TM1 plants silenced for *ILK* homologs

A) Box and whisker plot displaying RKN susceptibility as measured by RKN egg counts per gram of root six weeks after inoculation with 25,000 eggs. From left to right, egg count measurements are displayed for empty vector infiltrated plants, actin silenced plants, *GhILK1.2* silenced plants, *GhILK1.3* silenced plants, plants silenced for a combination of all *GhILK*s, and plants silenced for *GhILK1.1*. ** indicates significance at a p-value < 0.01. B) Assessment of target gene silencing 2 weeks after infiltration for plants infiltrated with a mixed inoculum targeting all *GhILK*s. C) Target gene silencing 2 weeks after infiltration for plants infiltrated with inoculum targeting a set of *GhILK*s. For B and C, only a single plant was taken to quantify silencing, mean expression values are plotted with error bars representing standard deviation within technical replicates. All values are relative to target gene expression in empty vector inoculated plant as normalized using a polyubiquitin housekeeping gene.

Verification of *GhILK1.1* in RKN resistance

The second round of silencing, examining a larger sampling of plants strongly silenced for *GhILK1.1* revealed a significant (p-value 0.004) increase in susceptibility to RKN in TM-1 when compared to empty vector controls (Figure 3.4a). No change in susceptibility to reniform nematodes was observed, although the removal of three potential outliers from the empty vector dataset suggests a potential increase in susceptibility to reniform nematodes as well (p-value 0.01). Further experiments are needed to explore and verify this observation.

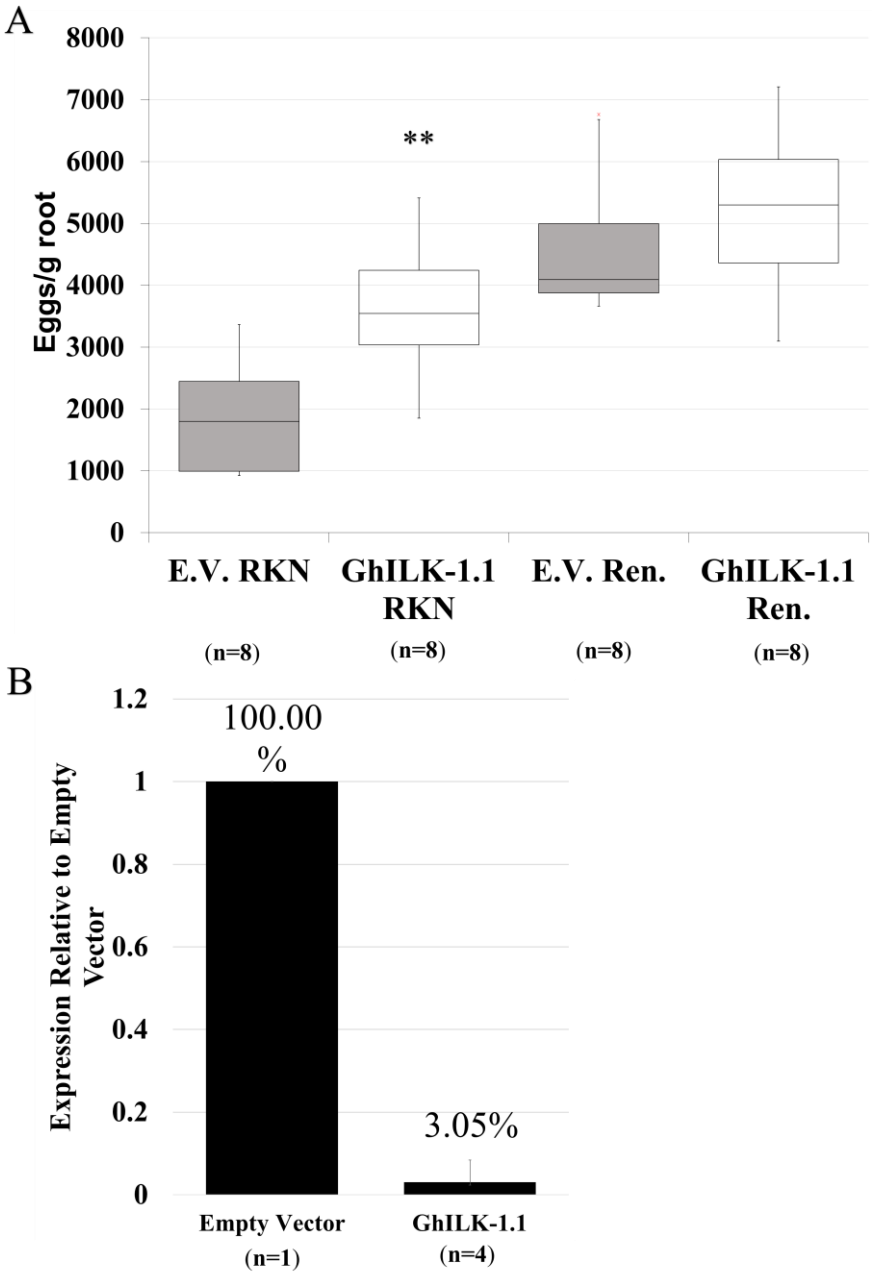


Figure 3.6 *GhILK1.1* functions in RKN and not reniform resistance

A) Box and whisker plot displaying RKN and reniform susceptibility as measured by egg counts per gram of root weight six weeks after inoculation with 35,000 eggs/pot for RKN and 14,000 eggs/pot for reniform nematodes. ** indicates significance at a p-value < 0.01. Red stars indicate outliers. B) Assessment of target gene silencing 2 weeks after inoculation. Values below treatment identifiers represent biological replicates. Error bars representative of the standard deviation between target gene expression in knockdown plants relative to empty vector control.

Discussion

The agro-drench and agro-infiltration methods worked with varying degrees of efficacy. Developed and tested within *Solanaceae*, the agro-drench method appears to have been less effective at inducing RNAi mediated gene silencing in cotton compared to the more conventional agro inoculation method. The modification of the geminivirus carrying the RNAi inducing fragment might also have contributed to the loss in silencing efficacy. Originally, the agro-drench method was performed using a Tobacco rattle virus (TRV)-derived VIGS silencing vector, whereas the current implementation utilized a ClCrV-derived silencing vector. Although both RNA and DNA virus have been developed to elicit VIGS in plants, TRV is an RNA virus while ClCrV is a DNA virus, perhaps contributing to the lack of efficacy from the agro-drench method (72).

Furthermore, a different strain of *Agrobacterium* (originally GV2260, we used GV3101) was used to carry the VIGS vector; previous studies have observed differences in target protein expression in plants transformed with different strains of *Agrobacterium* (73,74). Finally, previous reports have noted that the DNA abrasion method is also not an effective method of inoculating ClCrV-derived silencing vectors into cotton as both the mechanical properties of the host leaf and the specificity of the virus affect the silencing viability for all virus/host combinations (71). As the agro inoculation method resulted in easily distinguishable and significant phenotypic changes in cotton seedlings, it remains a consistently effective silencing method.

The silencing of *GhAct7* - previously characterized as having a moderate expression in both root and stem tissue, with 3-fold higher expression in roots (70) - resulted in a striking phenotype change and is an appropriate tool to visually confirm the

success of gene silencing in cotton root systems. Significant and easily distinguishable stunting in root and stem development were observed in inoculated cotton seedlings silenced for the *GhAct7* gene. This stunting of tissue development is due to a reduced capacity to produce actin proteins necessary for optimal cell elongation during early plant growth.

Previous studies have shown that RNAi mediated target gene silencing is strongly correlated with observed phenotypic changes in plants (75,76). Furthermore, mixing multiple *Agrobacterium* cultures for injection into a single plant has also been shown to significantly reduce both silencing efficiency and the observed magnitude of phenotypic changes in transiently silenced cotton plants (77). Although a 78% reduction in *GhILK1.1* resulted in a slight (albeit not significant) increase in susceptibility in the mixed *ILK* silenced seedlings, around a 90% knockdown of the *GhILK1.1* gene is needed for a significant increase in susceptibility to RKN. While the silencing of *GhILK1.1* was both successful and significantly correlated with an increase in RKN susceptibility, the silencing of *GhILK1.2* and *GhILk1.3* were less informative. Although *GhILK1.3* was shown to display an almost 70% decrease in expression compared to empty vector controls, it did not result in a significant change in susceptibility to RKN. Improvements to silencing efficiency could be leveraged by optimizing silencing constructs through modification of the insert fragment used for targeted gene silencing. Further improvements might also be available with alternative silencing methods, perhaps using Crisper/Cas9 mediated genome editing techniques.

The recently sequenced allotetraploid genome of *Gossypium hirsutum* is complex and abundant with presently uncharacterized gene duplicates. Recent whole genome

duplications appear to have contributed to the majority of gene duplicates within the MAP3K gene family (78). Further, genes involved in signal transduction and stress responses – core functions of plant MAP3Ks – have been shown to be preferentially retained following repeated rounds of WGD and tend to have numerous paralogs (36). This retention and gene family expansion is clearly evident in *G. hirsutum* MAP3Ks as its recent ploidy events have resulted in significant expansions within the *ILK* subfamily. Functional characterization of *GhILK1.1* revealed conserved pathogen immune responses predicted from its structural and phylogenetic homology to the Arabidopsis *ILK1* gene. Further work is needed to characterize *GhILK1.2* and *GhILK1.3*, which might function similarly in RKN resistance once adequately silenced. Alternatively, these paralogs may have neofunctionalized to mediate resistance to different pathogens entirely, or perhaps pseudogenized into functionally inactive kinases.

It has recently been shown by our lab that the Arabidopsis *ILK1* gene is involved in flg22 responses and resistance to bacterial pathogens. More specifically, *ILK1* is coupled to cellular K^+ fluxes through interactions with *HAK5* and is an interactor with the Ca^+ sensing *CML9* (9). These K^+ fluxes are known in animals to initiate defense pathways against intracellular recognition of PAMPs (79); likewise, in Arabidopsis, K^+ fluxes mediated by *ILK1* were also required for PAMP-triggered PM depolarization and subsequence signaling cascades involved in basal plant immunity. These changes in intracellular conditions have also been shown to activate MAPK signaling cascades and regulate downstream transcriptional reprogramming through regulation of early *MPK3*/*MPK6* signaling. We hypothesize that *G. hirsutum* homologs of the Arabidopsis *ILK1* retain similar functions in PTI during plant-nematode interactions. Although not

presently explored, similar signaling responses might underlie *GhILK1.1* mediated RKN resistance. An examination of PTI-induced ROS generation dynamics might be informative as studies in Arabidopsis have revealed *ILK1* does not appear to play a role in altering the generation of ROS. Further, an examination of orthologous *MPK3*/*MPK6* expression in *GhILK1.1* compromised cotton would verify an equivalent signaling pathway between *ILK1*-mediated PTI responses in both Arabidopsis and upland cotton. Finally, a comparison between *GhILK1.1*-deficient and control cotton root cells might uncover interesting dynamics related to potential early *ILK1*-mediated nutrient diversion by invading RKN during giant cell formation.

Conclusion

This study produced a gene characterization assay able to transiently silence and characterized sets of homeologs in upland cotton. *GhAct7* was shown to produce easily distinguishable phenotype changes compared to empty vector controls, and can be used to visually assess transient gene silencing in vivo.

GhILK1 is the first Raf-like MAP3K shown to be involved in providing resistance against RKN in cotton. Six total cotton homologs of the Arabidopsis *ILK1* gene were identified in the present study, consisting of 3 sets of almost identical homeologs. The *ILK* subfamily appears to have expanded from 6 members in Arabidopsis to 20 members in upland cotton, with the largest expansion observed for cotton homologs of *ILK6*/At1G14000. Previous characterization of the *ILK1* gene in Arabidopsis suggests a role of *GhILK1.1* in basal immune responses against a broad multitude of plant pathogens. Although *GhILK1.1* silencing significantly increased susceptibility to RKN in

the previously susceptible TM1 line, no significant change in susceptibility was noticed in TM1 against reniform nematodes.

Further, no significant change in resistance was observed following silencing of *GhILK1.1* in the RKN resistant line M240 against either RKN or reniform nematodes. *GhILK1.1* paralogs did not appear to be efficiently silenced and might have neo/subfunctionalized or pseudogenized to respond to different plant pathogens or lose functionality altogether. Further work is needed to confirm these predictions and assess paralog functionality.

Methods

Orthogroup identification

All gene models for the complete proteomes of *Arabidopsis thaliana* (80), *Gossypium raimondii* (49), *Gossypium hirsutum* (4), *Solanum lycopersicum* (50), *Glycine max* (51), *Zostera marina* (6), and *Zea mays* (52) were retrieved from Phytozome v12 (5). Sequence information was uploaded onto the Cyverse Discovery environment and clustered into OrthoMCL defined orthogroups using the workflow detailed at https://pods.iplantcollaborative.org/wiki/pages/viewpage.action?pageId=12881253. The all-vs-all BLASTp was conducted using the default E-value cutoff while returning up to the top 300 hits and alignments for each query for input into the OrthoMCL pipeline (81). The "OrthoMCL v1.4" application was used to cluster orthologs with an index mode set to all, a p-value cutoff of 1.5, percent identity cutoff of 0, a maximum weight of 350, and an inflation parameter of 1.5. Re-annotated orthogroups were then queried with the Arabidopsis *ILK1* protein in order to identify the *ILK1* subfamily in all seven species.

Sequence analysis

Sequences were retrieved for all members of the *ILK1* orthocluster and aligned in MEGA v7.0.26 (56) using MUSCLE (55). The maximum likelihood tree was built using the JTT + G model with support for nodes generated using 500 bootstrap replicates. Motif domains were identified using the EMBL SMART search tool(82) and NCBI's Conserved Domain Database search tool (31). Three orthoclusters contained all homologs of all six Arabidopsis *ILK*s; more clusters containing additional genes unique to cotton or absent from Arabidopsis are possible.

VIGS construction

Fragments for VIGS silencing of *G. hirsutum ILK*s were generated by identifying regions within coding sequences unique to a set of cotton homeologs. Homeolog sequences were examined using NCBI's CDD to avoid regions within broadly conserved functional domains. VIGS target regions were checked using the SGN VIGS tool (83) for possible off-target silencing. Using an n-mer size of 21 and not allowing for any mismatches, all three *GhILK1* silencing fragments returned zero potential off-target hits against the NBI v1.1 background databases along the entirety of the insert fragment. Primers flanking these regions were designed and modified to include HpaI and SpeI restriction sites allowing for insert of target silencing fragments into the ClCrV plasmid in an antisense manner. Insert fragments were amplified from cotton seedling root DNA extracted using a Qiagen DNeasy Plant mini kit using the primers listed in Appendix B.1-3. HpaI and SpeI restriction enzymes were used to digest the amplified fragments before ligation into the previously characterized pJRT.Agro.ClCrVa.008 plasmid(71). Ligates were transformed into e. coli DH5α competent cells, grown overnight on LB plates

containing 50 µg/ml kanamycin, and resulting colonies were checked for successful incorporation of the silencing fragment using colony PCR by amplifying region around the insert site and checking for the presence of the insert fragment at the predicted length. Plasmids were extracted from successfully transformed e. coli and transformed into competent *Agrobacterium* GV3101 cells using the heat-shock method. Transformed *Agrobacterium* cells were grown at 28°C for 2 days on LB plates containing 25 µg/ml gentamycin, kanamycin, and rifampicin. *Agrobacterium* was subsequently checked with colony PCR by using insert fragment specific primers. Glycerol stocks were prepared from successfully transformed *Agrobacterium* and used to prepare future inoculums.

RNA extraction and qRT-PCR

RNA was extracted from 100mg of root tissue using the protocol provided in Spectrum Plant Total RNA extraction kits. cDNA was generated by following the provided protocol in the iScript gDNA Clear cDNA Synthesis Kit from Biorad. qRT-PCR was performed on an ABI StepOnePlus instrument in 20ul reactions using SsoAdvanced Universal SYBR Green supermix.

Plant material and growth conditions

TM1 and M240 cotton seeds were scarified by gently filing the seed coat and allowed to pre-germinate in wet paper towels for around 8 hours. For the examination of *GhAct7*, individual germinated seeds were planted in Cone-tainers filled with a mixture of autoclaved sand and loam and fertilized with a small amount of Osmocote. Plants were watered every other day with purified water and maintained in growth chambers at 22°C with a 16h/8h day/night cycle. For later experiments examining *ILK1* homologs in

cotton, to reduce variance between individual measurements, 2 plants were grouped in small pots and used to assess RKN or reniform susceptibility following gene knockdown.

Transient gene silencing

The inoculum was prepared using glycerol stocks of previously described silencing constructs. From glycerol stocks, previously transformed *Agrobacterium* was grown at 28°C for two days on LB plates containing 25 µg/mL gentamycin, kanamycin, and rifampicin. Cells were harvested from agar plates, and a 5 mL culture was grown overnight in a 28°C shaker at 180 RPM in LB containing 25 µg/mL gentamycin and kanamycin. The next day, the overnight culture was used to inoculate a 50 mL flask of LB containing antibiotics (25 µg/mL gentamycin, kanamycin, and rifampicin), 10mM MES, and 20 uM acetosyringone and allowed to grow overnight in a 28°C shaker at 180 RPM. The next morning, cultures were centrifuged at 4000xg for 10 minutes at 4°C, and the bacterial pellet was resuspended in inoculation buffer containing 10mM $MgCl_2$, 10mM MES, and 200uM acetosyringone. Cultures were adjusted to an OD_{600} of 1.5 and left at room temperature for 3 to 4 hours before inoculation. A and B component were mixed immediately before either agro-drench or agro-infiltration in a 1:1 ratio of total A to total B components. For the agro-drench experiments, 1 week old cotton seedlings were drenched with 5 mL of inoculum, as recommended in the original protocol (73). For agro-infiltration, 1 week old seedlings had the underside of cotyledons perforated with a needle and inoculum was injected into the entirety of both cotyledons using a needleless 5 mL syringe. Following either infiltration or agro-drench, seedlings were covered with plastic and allowed to co-inoculate overnight at room temperature before being uncovered and returned to growth chambers.

Nematode assay

Two weeks after *Agrobacterium*-mediated silencing, seedlings were inoculated with nematodes. For the first *ILK* examination experiment, the soil subsurface was inoculated with 25,000 RKN eggs/pot at three distinct spots within pots. For the second experiment, to increase final egg counts, 35,000 eggs/pot were inoculated for RKN, and 14,000 eggs/pot were inoculated for reniform nematode assays. Seedlings were allowed to continue developing for 6 more weeks before roots were examined for nematode susceptibility (measured as eggs/g root weight). Root tissue was washed of soil mixture and dried before fresh root weight was measured. Eggs were extracted by soaking roots in 1% sodium hypochlorite for 3 minutes with agitation. The resulting solution was poured over a #200/#500 standard sieve stack where the eggs collect on the #500 sieve. The eggs were transferred with 50-80 ml water to a 120 ml sample cup and stained with acid fuchsin. All eggs within 1 ml were counted using a stereo-dissecting microscope. Total egg counts were then extrapolated for individual pots.

CHAPTER IV

CONCLUSION

MAPK signaling cascades have previously been identified and characterized in many plant species, primarily due to their roles in regulating plant growth, development, and responses to stress. MAP3Ks represent the largest, most structurally diverse gene family within MAPK cascades and function as the principal regulators of a diverse array of downstream signal transduction events. Although great care has previously been taken to fully identify the expanded MAP3K gene family in multiple plants, superior search algorithms and the recent increase in the number of well-sequenced plant genomes have made it possible to refine the MAP3K gene family in angiosperms using our newly proposed comparative gene family identification algorithm.

By leveraging well-annotated reference gene families to identify equivalent gene families in new species, we've refined the architecture of the MAP3K gene family in seven plant species. Previous annotations of MAP3Ks within five plants were in strong agreement with our newly proposed MAP3K gene families, although significant expansions within certain subfamilies were identified. Further, the MAP3K gene families within *G. hirsutum* and *Z. marina* were identified for the first time. Although a more extensive examination is required before functionality can be associated with proposed clades, transcriptome information was used to annotate subfamily clades with

putative conserved functionality. Gene duplication analyses was undertaken to examine the origins of presently identified plant MAP3Ks. Finally, gene collinearity was examined to define how the cotton MAP3K gene family has expanded so drastically compared to other angiosperms.

The novel identification of MAP3Ks within *G. hirsutum* was then leveraged to functionally characterize an ortholog of the previously examined Arabidopsis Raf-like MAP3K *ILK1*. The *ILK* subfamily was shown to have expanded significantly within both *G. hirsutum* and *G. raimondii* compared to Arabidopsis, and six *G. hirsutum* homologs (3 sets of homeologs) were selected for functional characterization. Although more work needs to be done to accurately assess the functionality of 2 sets of homeologs, transient gene silencing using newly created VIGS constructs revealed that a set of homeologs – presently identified as *GhILK1.1* - functions in mediating RKN resistance in upland cotton. *GhILK1.1* was further shown to have no significant effect on reniform nematode resistance. *GhILK1.1* represents the first cotton MAP3K shown to be involved in mediating cotton resistance to RKN.

In summary, a new method for comparative gene family classification was demonstrated – identifying for the first time the MAP3K gene families in cotton and seagrass – before a subset of cotton MAP3Ks were characterized and demonstrated to have critical contributions to RKN resistance in upland cotton.

REFERENCES

1. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, et al. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res. 2001 Jan 1;29(1):102–5.

2. Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, et al. The Sequenced Angiosperm Genomes and Genome Databases. Front Plant Sci [Internet]. 2018 [cited 2018 Sep 13];9. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2018.00418/full

3. Veeckman E, Ruttink T, Vandepoele K. Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. The Plant Cell. 2016 Aug 1;28(8):1759–68.

4. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, et al. Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nat Biotech. 2015 May;33(5):531–7.

5. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012 Jan;40(Database issue):D1178–86.

6. Olsen JL, Rouzé P, Verhelst B, Lin Y-C, Bayer T, Collen J, et al. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. Nature. 2016 Feb;530(7590):331–5.

7. Meng X, Zhang S. MAPK Cascades in Plant Disease Resistance Signaling. Annual Review of Phytopathology. 2013 Aug 4;51(1):245–66.

8. Melotto M, Panchal S, Roy D. Plant innate immunity against human bacterial pathogens. Front Microbiol [Internet]. 2014 Aug 11;5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4127659/

9. Brauer EK, Ahsan N, Dale R, Kato N, Coluccio AE, Piñeros MA, et al. The Raf-like Kinase ILK1 and the High Affinity K+ Transporter HAK5 Are Required for Innate Immunity and Abiotic Stress Response1[OPEN]. Plant Physiol. 2016 Jun;171(2):1470–84.

10. Holbein J, Grundler FMW, Siddique S. Plant basal resistance to nematodes: an update. J Exp Bot. 2016 Mar 1;67(7):2049–61.

11. Chinchilla D, Zipfel C, Robatzek S, Kemmerling B, Nürnberger T, Jones JDG, et al. A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence. Nature. 2007 Jul;448(7152):497–500.

12. Teixeira MA, Wei L, Kaloshian I. Root-knot nematodes induce pattern-triggered immunity in Arabidopsis thaliana roots. New Phytologist. 2016 Jul 1;211(1):276–87.

13. Wubben MJ, Callahan FE, Velten J, Burke JJ, Jenkins JN. Overexpression of MIC-3 indicates a direct role for the MIC gene family in mediating Upland cotton (Gossypium hirsutum) resistance to root-knot nematode (Meloidogyne incognita). Theor Appl Genet. 2015 Feb 1;128(2):199–209.

14. Zulawski M, Schulze G, Braginets R, Hartmann S, Schulze WX. The Arabidopsis Kinome: phylogeny and evolutionary insights into functional diversification. BMC Genomics. 2014 Jul 1;15:548.

15. Zhou H, Ren S, Han Y, Zhang Q, Qin L, Xing Y. Identification and Analysis of Mitogen-Activated Protein Kinase (MAPK) Cascades in Fragaria vesca. Int J Mol Sci [Internet]. 2017 Aug 13;18(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5578155/

16. Kong X, Lv W, Zhang D, Jiang S, Zhang S, Li D. Genome-wide identification and analysis of expression profiles of maize mitogen-activated protein kinase kinase kinase. PLoS ONE. 2013;8(2):e57714.

17. Sun Y, Wang C, Yang B, Wu F, Hao X, Liang W, et al. Identification and functional analysis of mitogen-activated protein kinase kinase kinase (MAPKKK) genes in canola (Brassica napus L.). J Exp Bot. 2014 May 1;65(8):2171–88.

18. Yin Z, Wang J, Wang D, Fan W, Wang S, Ye W. The MAPKKK Gene Family in Gossypium raimondii: Genome-Wide Identification, Classification and Expression Analysis. Int J Mol Sci. 2013 Sep 11;14(9):18740–57.

19. Rao KP, Richa T, Kumar K, Raghuram B, Sinha AK. In Silico Analysis Reveals 75 Members of Mitogen-Activated Protein Kinase Kinase Kinase Gene Family in Rice. DNA Research. 2010 Jun 1;17(3):139–53.

20. Li W, Xu H, Liu Y, Song L, Guo C, Shu Y. Bioinformatics Analysis of MAPKKK Family Genes in Medicago truncatula. Genes. 2016 Apr 4;7(4):13.

21. Wu J, Wang J, Pan C, Guan X, Wang Y, Liu S, et al. Genome-Wide Identification of MAPKK and MAPKKK Gene Families in Tomato and Transcriptional Profiling Analysis during Development and Stress Response. PLOS ONE. 2014 Jul 18;9(7):e103032.

22. Neupane A, Nepal MP, Piya S, Subramanian S, Rohila JS, Reese RN, et al. Identification, nomenclature, and evolutionary relationships of mitogen-activated protein kinase (MAPK) genes in soybean. Evol Bioinform Online. 2013;9:363–86.

23. Wang G, Lovato A, Polverari A, Wang M, Liang Y-H, Ma Y-C, et al. Genome-wide identification and analysis of mitogen activated protein kinase kinase kinase gene family in grapevine (Vitis vinifera). BMC Plant Biology [Internet]. 2014 Dec [cited 2018 Aug 3];14(1). Available from: http://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-014-0219-1

24. MAPK Group. Mitogen-activated protein kinase cascades in plants: a new nomenclature. Trends Plant Sci. 2002 Jul;7(7):301–8.

25. Lehti-Shiu MD, Shiu S-H. Diversity, classification and function of the plant protein kinase superfamily. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012 Sep 19;367(1602):2619–39.

26. Martinez M. Plant protein-coding gene families: emerging bioinformatics approaches. Trends in Plant Science. 2011 Oct 1;16(10):558–67.

27. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Research. 2017 Jan 4;45(D1):D183–9.

28. Hanks SK, Hunter T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J. 1995 May;9(8):576–96.

29. Li J, Mahajan A, Tsai M-D. Ankyrin Repeat: A Unique Motif Mediating Protein−Protein Interactions. Biochemistry. 2006 Dec 1;45(51):15168–78.

30. Zhao C, Nie H, Shen Q, Zhang S, Lukowitz W, Tang D. EDR1 Physically Interacts with MKK4/MKK5 and Negatively Regulates a MAP Kinase Cascade to Modulate Plant Innate Immunity. PLOS Genetics. 2014 May 15;10(5):e1004389.

31. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017 Jan 4;45(D1):D200–3.

32. Blanc G, Wolfe KH. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. Plant Cell. 2004 Jul;16(7):1667–78.

33. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, et al. The draft genome of a diploid cotton Gossypium raimondii. Nat Genet. 2012 Oct;44(10):1098–103.

34. Popescu SC, Popescu GV, Bachan S, Zhang Z, Gerstein M, Snyder M, et al. MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays. Genes Dev. 2009 Jan 1;23(1):80–92.

35. Rodriguez M, Petersen M, Mundy J. Mitogen-Activated Protein Kinase Signaling in Plants. Annual Review of Plant Biology. 2010 Jun 2;61(1):621–49.

36. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of Gene Duplication in Plants1[OPEN]. Plant Physiol. 2016 Aug;171(4):2294–316.

37. Zhu T, Liang C, Meng Z, Sun G, Meng Z, Guo S, et al. CottonFGD: an integrated functional genomics database for cotton. BMC Plant Biol [Internet]. 2017 Jun 8;17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5465443/

38. Lee B, Henderson DA, Zhu J-K. The Arabidopsis Cold-Responsive Transcriptome and Its Regulation by ICE1. The Plant Cell. 2005 Nov 1;17(11):3155–75.

39. Ma S, Bohnert HJ. Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression. Genome Biology. 2007 Apr 4;8:R49.

40. Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, Mittler R. When Defense Pathways Collide. The Response of Arabidopsis to a Combination of Drought and Heat Stress. Plant Physiology. 2004 Apr 1;134(4):1683–96.

41. Kim J-M, Woo D-H, Kim S-H, Lee S-Y, Park H-Y, Seok H-Y, et al. Arabidopsis MKKK20 is involved in osmotic stress response via regulation of MPK6 activity. Plant Cell Rep. 2012 Jan 1;31(1):217–24.

42. Jia H, Hao L, Guo X, Liu S, Yan Y, Guo X. A Raf-like MAPKKK gene, GhRaf19, negatively regulates tolerance to drought and salt and positively regulates resistance to cold stress by modulating reactive oxygen species in cotton. Plant Science. 2016 Nov 1;252:267–81.

43. Hashimoto M, Negi J, Young J, Israelsson M, Schroeder JI, Iba K. Arabidopsis HT1 kinase controls stomatal movements in response to CO2. Nat Cell Biol. 2006 Apr;8(4):391–7.

44. Hayashi M, Inoue S, Ueno Y, Kinoshita T. A Raf-like protein kinase BHP mediates blue light-dependent stomatal opening. Scientific Reports. 2017 Mar 30;7:45586.

45. Lamberti G, Gügel IL, Meurer J, Soll J, Schwenkert S. The Cytosolic Kinases STY8, STY17, and STY46 Are Involved in Chloroplast Differentiation in Arabidopsis. Plant Physiology. 2011 Sep 1;157(1):70–85.

46. Na Zhai, Haihong Jia, Dongdong Liu, Shuchang Liu, Manli Ma, Xingqi Guo, et al. GhMAP3K65, a Cotton Raf-Like MAP3K Gene, Enhances Susceptibility to Pathogen Infection and Heat Stress by Negatively Modulating Growth and Development in Transgenic Nicotiana benthamiana. International Journal of Molecular Sciences. 2017 Nov 21;18(11):2462.

47. Chen X, Wang J, Zhu M, Jia H, Liu D, Hao L, et al. A cotton Raf-like MAP3K gene, GhMAP3K40, mediates reduced tolerance to biotic and abiotic stress in Nicotiana benthamiana by negatively regulating growth and development. Plant Science. 2015 Nov 1;240:10–24.

48. Cheng C-Y, Krishnakumar V, Chan A, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. bioRxiv. 2016 Apr 5;47308.

49. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 2012 Dec;492(7429):423–7.

50. Consortium TTG. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012 May;485(7400):635–41.

51. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010 Jan;463(7278):178–83.

52. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science. 2009 Nov 20;326(5956):1112–5.

53. Li L. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research. 2003 Sep 1;13(9):2178–89.

54. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 2009 Jul 1;37(suppl_2):W202–8.

55. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

56. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016 Jul 1;33(7):1870–4.

57. He Z, Zhang H, Gao S, Lercher MJ, Chen W-H, Hu S. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. 2016 Jul 8;44(Web Server issue):W236–41.

58. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012 Apr;40(7):e49.

59. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res [Internet]. 2009 Jun 18 [cited 2018 May 21]; Available from: http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109

60. Jiang C-H, Xie P, Li K, Xie Y-S, Chen L-J, Wang J-S, et al. Evaluation of root-knot nematode disease control and plant growth promotion potential of biofertilizer Ning shield on Trichosanthes kirilowii in the field. Brazilian Journal of Microbiology. 2018 Apr 1;49(2):232–9.

61. Wheeler TA, Siders KT, Anderson MG, Russell SA, Woodward JE, Mullinix BG. Management of Meloidogyne incognita with Chemicals and Cultivars in Cotton in a Semi-Arid Environment. J Nematol. 2014 Jun;46(2):101–7.

62. Daudi A, Sawadogo A, Mateille T, Trivino C, Netscher C, Fargette M, et al. The importance of tropical root-knot nematodes (Meloidogyne spp.) and factors affecting the utility of Pasteuria penetrans as a biocontrol agent. Nematology. 2000 Nov 1;2(8):823–45.

63. Starr JL, Koenning SR, Kirkpatrick TL, Robinson AF, Roberts PA, Nichols RL. The Future of Nematode Management in Cotton. J Nematol. 2007 Dec;39(4):283–94.

64. Yadav U. RECENT TRENDS IN NEMATODE MANAGEMENT PRACTICES: THE INDIAN CONTEXT. International Research Journal of Engineering and Technology (IRJET). 2017;

65. Mitkowski NA, Abawi GS. Root-knot nematodes. The Plant Health Instructor [Internet]. 2003 [cited 2018 Sep 17]; Available from: http://www.apsnet.org/edcenter/intropp/lessons/Nematodes/Pages/RootknotNematode.aspx

66. Abad P, Gouzy J, Aury J-M, Castagnone-Sereno P, Danchin EGJ, Deleury E, et al. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nature Biotechnology. 2008 Aug;26(8):909–15.

67. Mitchum MG, Hussey RS, Baum TJ, Wang X, Elling AA, Wubben M, et al. Nematode effector proteins: an emerging paradigm of parasitism. New Phytologist. 2013 Sep 1;199(4):879–94.

68. Cristina M, Petersen M, Mundy J. Mitogen-Activated Protein Kinase Signaling in Plants. Annual Review of Plant Biology. 2010 Jun 2;61(1):621–49.

69. Popescu SC, Brauer EK, Dimlioglu G, Popescu GV. Insights into the Structure, Function, and Ion-Mediated Signaling Pathways Transduced by Plant Integrin-Linked Kinases. Front Plant Sci [Internet]. 2017 [cited 2017 Apr 24];8. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2017.00376/full#B5

70. Li X-B. The Cotton ACTIN1 Gene Is Functionally Expressed in Fibers and Participates in Fiber Elongation. THE PLANT CELL ONLINE. 2005 Feb 18;17(3):859–75.

71. Tuttle J, Haigler CH, Robertson D. Method: low-cost delivery of the cotton leaf crumple virus-induced gene silencing system. Plant Methods. 2012;8(1):27.

72. Lange M, Yellina AL, Orashakova S, Becker A. Virus-Induced Gene Silencing (VIGS) in Plants: An Overview of Target Species and the Virus-Derived Vector Systems. In: Becker A, editor. Virus-Induced Gene Silencing [Internet]. Totowa, NJ: Humana Press; 2013 [cited 2018 Oct 23]. p. 1–14. Available from: http://link.springer.com/10.1007/978-1-62703-278-0_1

73. Ryu C-M, Anand A, Kang L, Mysore KS. Agrodrench: a novel and effective agroinoculation method for virus-induced gene silencing in roots and diverse Solanaceous species: RNA silencing by agrodrench. The Plant Journal. 2004 Sep 14;40(2):322–31.

74. Yadav S, Sharma P, Srivastava A, Desai P, Shrivastava N. Strain specific Agrobacterium-mediated genetic transformation of Bacopa monnieri. Journal of Genetic Engineering and Biotechnology. 2014 Dec 1;12(2):89–94.

75. Kanneganti V, Gupta AK. RNAi mediated silencing of a wall associated kinase, OsiWAK1 in Oryza sativa results in impaired root development and sterility due to anther indehiscence. Physiol Mol Biol Plants. 2011 Mar;17(1):65–77.

76. Travella S, Klimm TE, Keller B. RNA Interference-Based Gene Silencing as an Efficient Tool for Functional Genomics in Hexaploid Bread Wheat. Plant Physiol. 2006 Sep;142(1):6–20.

77. Pang J, Zhu Y, Li Q, Liu J, Tian Y, Liu Y, et al. Development of Agrobacterium-Mediated Virus-Induced Gene Silencing and Performance Evaluation of Four Marker Genes in Gossypium barbadense. PLOS ONE. 2013 Sep 2;8(9):e73211.

78. Bokros N, Popescu SC, Popescu GV. Multispecies genome-wide analysis defines the MAP3K gene family in Gossypium hirsutum and reveals conserved family expansions. BMC Bioinformatics (Accepted).

79. He Y, Zeng MY, Yang D, Motro B, Núñez G. NEK7 is an essential mediator of NLRP3 activation downstream of potassium efflux. Nature. 2016 Jan 27;530(7590):354–7.

80. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant Journal. 2017 Feb 1;89(4):789–804.

81. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 2003 Sep 1;13(9):2178–89.

82. Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. Nucleic Acids Research. 2015 Jan 28;43(D1):D257–60.

83. Fernandez-Pozo N, Rosli HG, Martin GB, Mueller LA. The SGN VIGS Tool: User-Friendly Software to Design Virus-Induced Gene Silencing (VIGS) Constructs for Functional Genomics. Molecular Plant. 2015 Mar 2;8(3):486–8.

84. Yang J, Moeinzadeh M-H, Kuhl H, Helmuth J, Xiao P, Haas S, et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nature Plants. 2017 Sep;3(9):696–703.

85. Al-Mohanna T, Hasan N, Bokros N, Dimlioglu G, Reddy R, Shankle M, et al. The leaf and root proteomes of sweet potato (Ipomoea batatas). Manuscript submitted for publication.

86. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nature Protocols. 2013 Aug;8(8):1551–66.

APPENDIX A

SUPPLEMENTS FOR CHAPTER II

Table A.1 All presently identified MAP3Ks in seven plant species

Gene ID	Length	Identity	Subfamily	Duplicate type
AT1G49160	604	Known	ZIK	WGD/segmental
AT1G64630	524	Known	ZIK	Dispersed
AT3G04910	700	Known	ZIK	WGD/segmental
AT3G18750	567	Known	ZIK	WGD/segmental
AT3G22420	666	Known	ZIK	Dispersed
AT3G48260	516	Known	ZIK	Dispersed
AT3G51630	549	Known	ZIK	Dispersed
AT5G28080	492	Known	ZIK	WGD/segmental
AT5G41990	563	Known	ZIK	Dispersed
AT5G55560	314	Known	ZIK	Dispersed
AT5G58350	571	Known	ZIK	Dispersed
Glyma.01G132000	609	Known	ZIK	WGD/segmental
Glyma.02G235700	608	Known	ZIK	WGD/segmental
Glyma.02G297200	300	Known	ZIK	WGD/segmental
Glyma.02G306500	297	Known	ZIK	WGD/segmental
Glyma.03G036500	610	Known	ZIK	WGD/segmental
Glyma.03G245500	734	Known	ZIK	WGD/segmental
Glyma.04G188800	599	New	ZIK	WGD/segmental
Glyma.06G151000	656	Known	ZIK	WGD/segmental
Glyma.06G176800	584	Known	ZIK	WGD/segmental
Glyma.07G053500	710	Known	ZIK	WGD/segmental
Glyma.08G325600	299	New	ZIK	WGD/segmental
Glyma.09G276300	618	Known	ZIK	WGD/segmental
Glyma.10G092400	226	New	ZIK	Dispersed
Glyma.10G160400	738	Known	ZIK	WGD/segmental
Glyma.10G248400	730	Known	ZIK	WGD/segmental
Glyma.11G159900	455	New	ZIK	WGD/segmental
Glyma.13G002100	618	Known	ZIK	WGD/segmental
Glyma.14G006600	297	Known	ZIK	WGD/segmental
Glyma.14G016400	299	Known	ZIK	WGD/segmental
Glyma.14G203700	607	Known	ZIK	WGD/segmental
Glyma.16G022600	674	Known	ZIK	WGD/segmental
Glyma.18G054100	610	Known	ZIK	WGD/segmental
Glyma.18G081500	299	Known	ZIK	WGD/segmental
Glyma.18G215100	540	Known	ZIK	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Glyma.19G242900	682	Known	ZIK	WGD/segmental
Glyma.20G066900	618	Known	ZIK	WGD/segmental
Glyma.20G145800	730	Known	ZIK	WGD/segmental
Glyma.20G228000	738	Known	ZIK	WGD/segmental
Gohir.1Z041900	631	New	ZIK	NA
Gohir.A01G086900	294	New	ZIK	WGD/segmental
Gohir.A02G031800	734	New	ZIK	WGD/segmental
Gohir.A02G097500	417	New	ZIK	WGD/segmental
Gohir.A05G386300	610	New	ZIK	WGD/segmental
Gohir.A07G056400	611	New	ZIK	WGD/segmental
Gohir.A08G124300	667	New	ZIK	WGD/segmental
Gohir.A09G244300	639	New	ZIK	WGD/segmental
Gohir.A10G149800	593	New	ZIK	WGD/segmental
Gohir.A11G150300	297	New	ZIK	WGD/segmental
Gohir.A11G226100	624	New	ZIK	WGD/segmental
Gohir.A11G255500	573	New	ZIK	WGD/segmental
Gohir.A12G239100	607	New	ZIK	WGD/segmental
Gohir.A13G235500	593	New	ZIK	WGD/segmental
Gohir.D01G073100	295	New	ZIK	WGD/segmental
Gohir.D02G038500	734	New	ZIK	WGD/segmental
Gohir.D04G023400	609	New	ZIK	WGD/segmental
Gohir.D07G060700	609	New	ZIK	WGD/segmental
Gohir.D08G145400	740	New	ZIK	WGD/segmental
Gohir.D09G245200	621	New	ZIK	WGD/segmental
Gohir.D10G117000	593	New	ZIK	WGD/segmental
Gohir.D11G156800	298	New	ZIK	WGD/segmental
Gohir.D11G231300	513	New	ZIK	WGD/segmental
Gohir.D11G265300	727	New	ZIK	WGD/segmental
Gohir.D12G240300	607	New	ZIK	WGD/segmental
Gohir.D13G139300	299	New	ZIK	WGD/segmental
Gohir.D13G241200	593	New	ZIK	WGD/segmental
Gorai.001G064400	600	New	ZIK	WGD/segmental
Gorai.002G096100	296	Known	ZIK	WGD/segmental
Gorai.003G069600	643	Known	ZIK	Dispersed
Gorai.003G075600	612	Known	ZIK	WGD/segmental
Gorai.004G150400	668	Known	ZIK	Dispersed
Gorai.005G042400	762	Known	ZIK	Dispersed
Gorai.006G269500	621	Known	ZIK	WGD/segmental
Gorai.007G167300	298	New	ZIK	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Gorai.007G285300	727	Known	ZIK	Dispersed
Gorai.008G257300	607	Known	ZIK	WGD/segmental
Gorai.011G128900	593	Known	ZIK	WGD/segmental
Gorai.012G033500	609	Known	ZIK	Dispersed
Gorai.013G156000	299	Known	ZIK	WGD/segmental
Gorai.013G272100	593	Known	ZIK	WGD/segmental
GRMZM2G021416	566	Known	ZIK	WGD/segmental
GRMZM2G023444	324	New	ZIK	WGD/segmental
GRMZM2G032619	667	New	ZIK	Dispersed
GRMZM2G034779	380	New	ZIK	WGD/segmental
GRMZM2G084791	565	Known	ZIK	WGD/segmental
GRMZM2G089159	703	Known	ZIK	Dispersed
GRMZM2G116376	451	Known	ZIK	Dispersed
GRMZM2G312970	592	Known	ZIK	WGD/segmental
GRMZM5G878530	610	Known	ZIK	Dispersed
Solyc01g096170.2	767	Known	ZIK	WGD/segmental
Solyc01g097840.2	748	Known	ZIK	Dispersed
Solyc03g112140.2	664	Known	ZIK	WGD/segmental
Solyc05g041420.2	362	Known	ZIK	Dispersed
Solyc06g071800.2	626	Known	ZIK	WGD/segmental
Solyc06g082470.2	636	Known	ZIK	Dispersed
Solyc07g047990.1	290	Known	ZIK	Dispersed
Solyc07g065250.2	304	Known	ZIK	Dispersed
Solyc08g082980.2	586	Known	ZIK	WGD/segmental
Solyc09g018170.2	606	Known	ZIK	Dispersed
Solyc09g076000.2	731	Known	ZIK	Dispersed
Solyc10g009060.1	322	Known	ZIK	Dispersed
Solyc10g009350.2	656	Known	ZIK	WGD/segmental
Zosma146g00540	719	New	ZIK	N/A
Zosma185g00500	650	New	ZIK	N/A
Zosma270g00120	554	New	ZIK	N/A
Zosma56g01260	647	New	ZIK	N/A
Zosma63g00150	641	New	ZIK	N/A
Zosma99g00620	292	New	ZIK	N/A
AT1G05100	339	Known	MEKK	WGD/segmental
AT1G07150	499	Known	MEKK	WGD/segmental
AT1G09000	666	Known	MEKK	WGD/segmental
AT1G53570	609	Known	MEKK	Dispersed

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
AT1G54960	651	Known	MEKK	WGD/segmental
AT1G63700	883	Known	MEKK	Dispersed
AT2G05060	315	Known	MEKK	Dispersed
AT2G30040	463	Known	MEKK	WGD/segmental
AT2G32510	372	Known	MEKK	WGD/segmental
AT2G34290	265	Known	MEKK	Dispersed
AT2G40500	295	Known	MEKK	Proximal
AT2G40560	303	Known	MEKK	Proximal
AT2G40580	311	Known	MEKK	Proximal
AT2G41910	373	Known	MEKK	Tandem
AT2G41920	318	Known	MEKK	Tandem
AT2G41930	351	Known	MEKK	Tandem
AT2G42550	344	Known	MEKK	Dispersed
AT3G06030	651	Known	MEKK	Dispersed
AT3G07980	1367	Known	MEKK	Dispersed
AT3G13530	1368	Known	MEKK	Dispersed
AT3G45670	379	Known	MEKK	Dispersed
AT3G45790	376	Known	MEKK	Dispersed
AT3G46140	376	Known	MEKK	Proximal
AT3G46160	393	Known	MEKK	Proximal
AT3G50310	342	Known	MEKK	WGD/segmental
AT4G08470	560	Known	MEKK	WGD/segmental
AT4G08480	773	Known	MEKK	Tandem
AT4G08500	608	Known	MEKK	Proximal
AT4G12020	1895	Known	MEKK	Dispersed
AT4G26890	444	Known	MEKK	WGD/segmental
AT4G36950	336	Known	MEKK	WGD/segmental
AT5G12090	369	Known	MEKK	Dispersed
AT5G27510	301	Known	MEKK	Dispersed
AT5G27790	327	Known	MEKK	Dispersed
AT5G55090	510	Known	MEKK	WGD/segmental
AT5G66850	716	Known	MEKK	Dispersed
AT5G67080	344	Known	MEKK	WGD/segmental
Glyma.01G043100	328	New	MEKK	Dispersed
Glyma.01G184000	633	Known	MEKK	WGD/segmental
Glyma.01G186900	346	New	MEKK	WGD/segmental
Glyma.01G220700	869	Known	MEKK	WGD/segmental
Glyma.02G228300	466	Known	MEKK	WGD/segmental
Glyma.03G106000	348	New	MEKK	Proximal

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Glyma.03G106200	346	New	MEKK	Proximal
Glyma.03G237600	662	Known	MEKK	WGD/segmental
Glyma.04G036300	655	Known	MEKK	WGD/segmental
Glyma.04G213000	601	New	MEKK	WGD/segmental
Glyma.04G253500	566	New	MEKK	WGD/segmental
Glyma.05G080900	634	Known	MEKK	WGD/segmental
Glyma.05G094400	341	New	MEKK	WGD/segmental
Glyma.05G123200	500	Known	MEKK	WGD/segmental
Glyma.05G191700	600	Known	MEKK	WGD/segmental
Glyma.06G036400	671	Known	MEKK	WGD/segmental
Glyma.06G108900	555	New	MEKK	WGD/segmental
Glyma.06G153200	616	Known	MEKK	WGD/segmental
Glyma.06G226000	300	New	MEKK	Dispersed
Glyma.06G243900	247	New	MEKK	Proximal
Glyma.06G244200	295	New	MEKK	Proximal
Glyma.06G307600	385	New	MEKK	WGD/segmental
Glyma.08G015500	1038	Known	MEKK	WGD/segmental
Glyma.08G078200	470	Known	MEKK	WGD/segmental
Glyma.08G156900	596	New	MEKK	WGD/segmental
Glyma.09G005500	422	Known	MEKK	WGD/segmental
Glyma.09G135100	897	Known	MEKK	WGD/segmental
Glyma.10G232800	887	Known	MEKK	WGD/segmental
Glyma.10G250800	624	New	MEKK	WGD/segmental
Glyma.11G022900	844	Known	MEKK	WGD/segmental
Glyma.11G054400	265	New	MEKK	Proximal
Glyma.11G055100	346	New	MEKK	WGD/segmental
Glyma.11G058300	623	Known	MEKK	WGD/segmental
Glyma.11G101700	1392	Known	MEKK	WGD/segmental
Glyma.11G170000	411	Known	MEKK	WGD/segmental
Glyma.12G027600	1380	Known	MEKK	WGD/segmental
Glyma.12G097200	352	Known	MEKK	WGD/segmental
Glyma.12G164900	329	Known	MEKK	WGD/segmental
Glyma.12G192500	396	Known	MEKK	WGD/segmental
Glyma.13G084100	594	New	MEKK	WGD/segmental
Glyma.13G309900	398	Known	MEKK	WGD/segmental
Glyma.14G080100	702	Known	MEKK	WGD/segmental
Glyma.14G165700	553	New	MEKK	WGD/segmental
Glyma.14G166000	590	New	MEKK	Proximal

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Glyma.14G195300	487	Known	MEKK	WGD/segmental
Glyma.15G048500	472	New	MEKK	WGD/segmental
Glyma.15G048600	440	Known	MEKK	Tandem
Glyma.16G001200	336	Known	MEKK	WGD/segmental
Glyma.16G181000	898	Known	MEKK	WGD/segmental
Glyma.17G173000	341	New	MEKK	WGD/segmental
Glyma.17G177900	637	Known	MEKK	WGD/segmental
Glyma.17G245300	568	Known	MEKK	WGD/segmental
Glyma.18G060900	387	Known	MEKK	WGD/segmental
Glyma.18G244200	290	New	MEKK	Dispersed
Glyma.19G235200	658	Known	MEKK	WGD/segmental
Glyma.20G142900	627	New	MEKK	WGD/segmental
Glyma.20G161500	888	Known	MEKK	WGD/segmental
Gohir.1Z039900	359	New	MEKK	NA
Gohir.A01G013000	1392	New	MEKK	WGD/segmental
Gohir.A01G082200	437	New	MEKK	WGD/segmental
Gohir.A02G068000	505	New	MEKK	WGD/segmental
Gohir.A02G074800	575	New	MEKK	WGD/segmental
Gohir.A02G156800	338	New	MEKK	WGD/segmental
Gohir.A03G028800	483	New	MEKK	WGD/segmental
Gohir.A03G164300	1446	New	MEKK	WGD/segmental
Gohir.A05G171900	451	New	MEKK	WGD/segmental
Gohir.A05G199900	531	New	MEKK	WGD/segmental
Gohir.A05G232100	590	New	MEKK	WGD/segmental
Gohir.A05G385300	661	New	MEKK	WGD/segmental
Gohir.A08G183300	722	New	MEKK	WGD/segmental
Gohir.A08G201200	519	New	MEKK	WGD/segmental
Gohir.A09G050200	896	New	MEKK	Dispersed
Gohir.A09G055300	667	New	MEKK	WGD/segmental
Gohir.A10G012200	446	New	MEKK	WGD/segmental
Gohir.A10G091500	336	New	MEKK	WGD/segmental
Gohir.A10G092900	609	New	MEKK	WGD/segmental
Gohir.A11G045900	711	New	MEKK	WGD/segmental
Gohir.A11G264700	663	New	MEKK	WGD/segmental
Gohir.A12G001400	659	New	MEKK	WGD/segmental
Gohir.A12G069900	691	New	MEKK	WGD/segmental
Gohir.A12G072100	350	New	MEKK	WGD/segmental
Gohir.A12G134300	1419	New	MEKK	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Gohir.A13G233100	480	New	MEKK	WGD/segmental
Gohir.D01G014300	1205	New	MEKK	WGD/segmental
Gohir.D01G068800	437	New	MEKK	WGD/segmental
Gohir.D02G074000	405	New	MEKK	WGD/segmental
Gohir.D02G081800	589	New	MEKK	WGD/segmental
Gohir.D02G187600	1364	New	MEKK	WGD/segmental
Gohir.D03G023900	338	New	MEKK	WGD/segmental
Gohir.D03G139800	466	New	MEKK	WGD/segmental
Gohir.D04G022500	662	New	MEKK	WGD/segmental
Gohir.D05G107500	461	New	MEKK	WGD/segmental
Gohir.D05G175000	407	New	MEKK	WGD/segmental
Gohir.D05G202900	617	New	MEKK	WGD/segmental
Gohir.D08G201900	897	New	MEKK	WGD/segmental
Gohir.D08G218400	469	New	MEKK	WGD/segmental
Gohir.D09G043100	896	New	MEKK	Dispersed
Gohir.D09G054200	666	New	MEKK	WGD/segmental
Gohir.D10G011600	446	New	MEKK	WGD/segmental
Gohir.D10G095400	572	New	MEKK	WGD/segmental
Gohir.D10G096700	336	New	MEKK	WGD/segmental
Gohir.D11G049500	711	New	MEKK	WGD/segmental
Gohir.D11G199800	359	New	MEKK	WGD/segmental
Gohir.D11G274900	663	New	MEKK	WGD/segmental
Gohir.D12G001200	658	New	MEKK	WGD/segmental
Gohir.D12G070000	350	New	MEKK	WGD/segmental
Gohir.D12G089300	724	New	MEKK	WGD/segmental
Gohir.D12G137900	1419	New	MEKK	WGD/segmental
Gohir.D13G238700	480	New	MEKK	WGD/segmental
Gorai.002G016000	1390	Known	MEKK	WGD/segmental
Gorai.002G091000	437	Known	MEKK	WGD/segmental
Gorai.003G025400	338	Known	MEKK	WGD/segmental
Gorai.003G146500	490	Known	MEKK	WGD/segmental
Gorai.004G213200	897	Known	MEKK	Dispersed
Gorai.004G232400	495	Known	MEKK	WGD/segmental
Gorai.005G083400	508	New	MEKK	WGD/segmental
Gorai.005G091700	589	Known	MEKK	WGD/segmental
Gorai.005G210100	1428	Known	MEKK	WGD/segmental
Gorai.006G049500	896	Known	MEKK	Dispersed
Gorai.006G052900	668	Known	MEKK	Dispersed

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Gorai.007G053300	711	Known	MEKK	WGD/segmental
Gorai.007G213900	359	New	MEKK	WGD/segmental
Gorai.007G296700	646	Known	MEKK	Dispersed
Gorai.008G000500	660	Known	MEKK	Dispersed
Gorai.008G073600	724	Known	MEKK	WGD/segmental
Gorai.008G077000	350	New	MEKK	WGD/segmental
Gorai.008G149400	1419	Known	MEKK	WGD/segmental
Gorai.009G111600	519	New	MEKK	WGD/segmental
Gorai.009G180300	526	Known	MEKK	WGD/segmental
Gorai.009G208800	747	New	MEKK	WGD/segmental
Gorai.009G242600	590	Known	MEKK	WGD/segmental
Gorai.011G012900	446	Known	MEKK	WGD/segmental
Gorai.011G104700	613	Known	MEKK	WGD/segmental
Gorai.011G106000	336	Known	MEKK	WGD/segmental
Gorai.012G034500	662	Known	MEKK	WGD/segmental
Gorai.013G269600	480	Known	MEKK	WGD/segmental
AC197029.3	284	New	MEKK	singleton
AC204050.4	470	New	MEKK	singleton
AC209208.3	988	Known	MEKK	singleton
GRMZM2G017654	1337	Known	MEKK	WGD/segmental
GRMZM2G034877	689	Known	MEKK	WGD/segmental
GRMZM2G041774	514	Known	MEKK	WGD/segmental
GRMZM2G044557	633	Known	MEKK	WGD/segmental
GRMZM2G064613	689	Known	MEKK	WGD/segmental
GRMZM2G066120	600	Known	MEKK	WGD/segmental
GRMZM2G093316	895	Known	MEKK	WGD/segmental
GRMZM2G098828	674	Known	MEKK	WGD/segmental
GRMZM2G130927	629	Known	MEKK	WGD/segmental
GRMZM2G140726	727	Known	MEKK	WGD/segmental
GRMZM2G156800	755	Known	MEKK	WGD/segmental
GRMZM2G158860	525	New	MEKK	Dispersed
GRMZM2G165099	475	Known	MEKK	WGD/segmental
GRMZM2G173965	472	Known	MEKK	WGD/segmental
GRMZM2G175504	887	Known	MEKK	WGD/segmental
GRMZM2G180555	599	Known	MEKK	WGD/segmental
GRMZM2G305066	479	Known	MEKK	WGD/segmental
GRMZM2G335826	375	New	MEKK	WGD/segmental
GRMZM2G378852	371	New	MEKK	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
GRMZM2G404078	434	New	MEKK	Tandem
GRMZM2G439350	456	Known	MEKK	WGD/segmental
GRMZM2G459824	490	New	MEKK	WGD/segmental
GRMZM2G476477	483	Known	MEKK	WGD/segmental
GRMZM2G540772	600	Known	MEKK	Dispersed
GRMZM6G513881	394	Known	MEKK	Dispersed
Solyc01g079750.2	688	Known	MEKK	WGD/segmental
Solyc01g098980.2	1618	Known	MEKK	WGD/segmental
Solyc01g103240.2	359	Known	MEKK	Dispersed
Solyc01g104530.2	665	Known	MEKK	Dispersed
Solyc02g064870.1	301	New	MEKK	Tandem
Solyc02g064930.1	318	Known	MEKK	Tandem
Solyc02g064980.1	359	Known	MEKK	WGD/segmental
Solyc02g065110.2	630	Known	MEKK	WGD/segmental
Solyc02g090430.2	638	Known	MEKK	WGD/segmental
Solyc02g090970.1	360	Known	MEKK	WGD/segmental
Solyc02g090980.1	355	Known	MEKK	Tandem
Solyc02g090990.1	356	Known	MEKK	Tandem
Solyc03g025360.2	890	Known	MEKK	WGD/segmental
Solyc03g117640.1	405	Known	MEKK	WGD/segmental
Solyc04g079400.2	715	Known	MEKK	WGD/segmental
Solyc06g036080.2	913	Known	MEKK	WGD/segmental
Solyc06g065660.1	325	New	MEKK	Proximal
Solyc06g065790.1	357	New	MEKK	Proximal
Solyc06g068510.1	426	Known	MEKK	WGD/segmental
Solyc07g047910.1	485	Known	MEKK	Dispersed
Solyc07g051860.1	326	Known	MEKK	WGD/segmental
Solyc07g051870.1	329	Known	MEKK	Tandem
Solyc07g051880.1	326	Known	MEKK	Tandem
Solyc07g051890.1	329	Known	MEKK	Tandem
Solyc07g051920.1	322	Known	MEKK	Tandem
Solyc07g051930.1	370	Known	MEKK	Tandem
Solyc07g053170.2	601	Known	MEKK	Dispersed
Solyc07g064820.1	490	Known	MEKK	Dispersed
Solyc08g069090.1	320	Known	MEKK	Proximal
Solyc08g076490.2	377	Known	MEKK	Dispersed
Solyc08g081210.2	840	Known	MEKK	Dispersed
Solyc11g006000.1	614	Known	MEKK	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Solyc11g033270.1	1401	New	MEKK	Dispersed
Solyc12g088940.1	680	Known	MEKK	WGD/segmental
Zosma100g00030	1417	New	MEKK	N/A
Zosma12g00430	1258	New	MEKK	N/A
Zosma135g00450	521	New	MEKK	N/A
Zosma164g00170	567	New	MEKK	N/A
Zosma16g00390	874	New	MEKK	N/A
Zosma193g00020	364	New	MEKK	N/A
Zosma209g00210	353	New	MEKK	N/A
Zosma22g00900	655	New	MEKK	N/A
Zosma253g00060	871	New	MEKK	N/A
Zosma26g00350	777	New	MEKK	N/A
Zosma334g00080	690	New	MEKK	N/A
Zosma388g00040	612	New	MEKK	N/A
Zosma60g00470	634	New	MEKK	N/A
Zosma64g00490	414	New	MEKK	N/A
Zosma68g00400	610	New	MEKK	N/A
Zosma70g00360	386	New	MEKK	N/A
Zosma88g00540	349	New	MEKK	N/A
Zosma88g00560	357	New	MEKK	N/A
Zosma91g00370	591	New	MEKK	N/A
Zosma96g00760	701	New	MEKK	N/A
AT1G04700	1042	Known	RAF	Dispersed
AT1G08720	933	Known	RAF	Dispersed
AT1G14000	438	Known	RAF	Dispersed
AT1G16270	1147	Known	RAF	Dispersed
AT1G18160	992	Known	RAF	Dispersed
AT1G62400	390	Known	RAF	Dispersed
AT1G67890	765	Known	RAF	Dispersed
AT1G73660	1030	Known	RAF	Dispersed
AT1G79570	1248	Known	RAF	Dispersed
AT2G17700	546	Known	RAF	Dispersed
AT2G24360	411	Known	RAF	Dispersed
AT2G31010	775	Known	RAF	Dispersed
AT2G31800	476	Known	RAF	Dispersed
AT2G35050	1257	Known	RAF	Dispersed
AT2G42640	781	Known	RAF	Dispersed
AT2G43850	479	Known	RAF	Dispersed

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
AT3G01490	411	Known	RAF	Dispersed
AT3G06620	773	Known	RAF	Dispersed
AT3G06630	698	Known	RAF	Dispersed
AT3G06640	730	Known	RAF	Dispersed
AT3G22750	378	Known	RAF	Dispersed
AT3G24715	1117	Known	RAF	Dispersed
AT3G27560	356	Known	RAF	Dispersed
AT3G46920	1155	Known	RAF	Tandem
AT3G46930	515	Known	RAF	Tandem
AT3G50720	377	Known	RAF	Tandem
AT3G50730	371	Known	RAF	WGD/segmental
AT3G58640	809	Known	RAF	WGD/segmental
AT3G58760	534	Known	RAF	WGD/segmental
AT3G59830	477	Known	RAF	WGD/segmental
AT3G63260	391	Known	RAF	WGD/segmental
AT4G14780	364	Known	RAF	WGD/segmental
AT4G18950	459	Known	RAF	WGD/segmental
AT4G23050	736	Known	RAF	WGD/segmental
AT4G24480	956	Known	RAF	WGD/segmental
AT4G31170	412	Known	RAF	WGD/segmental
AT4G35780	570	Known	RAF	WGD/segmental
AT4G38470	575	Known	RAF	WGD/segmental
AT5G01850	357	Known	RAF	WGD/segmental
AT5G03730	821	Known	RAF	WGD/segmental
AT5G07140	583	Known	RAF	WGD/segmental
AT5G11850	880	Known	RAF	WGD/segmental
AT5G40540	353	Known	RAF	WGD/segmental
AT5G49470	831	Known	RAF	WGD/segmental
AT5G50000	385	Known	RAF	WGD/segmental
AT5G50180	346	Known	RAF	WGD/segmental
AT5G57610	1054	Known	RAF	WGD/segmental
AT5G58950	525	Known	RAF	WGD/segmental
AT5G66710	405	Known	RAF	WGD/segmental
Glyma.01G052600	427	Known	RAF	WGD/segmental
Glyma.01G132300	371	Known	RAF	WGD/segmental
Glyma.01G161600	571	Known	RAF	WGD/segmental
Glyma.01G217100	781	Known	RAF	Dispersed
Glyma.01G236100	387	Known	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Glyma.02G165800	660	Known	RAF	Dispersed
Glyma.02G288300	454	Known	RAF	WGD/segmental
Glyma.03G036000	371	Known	RAF	WGD/segmental
Glyma.03G191000	810	Known	RAF	WGD/segmental
Glyma.04G020100	532	Known	RAF	WGD/segmental
Glyma.04G096000	927	Known	RAF	WGD/segmental
Glyma.04G181500	357	Known	RAF	WGD/segmental
Glyma.04G182700	386	New	RAF	WGD/segmental
Glyma.04G188200	352	Known	RAF	WGD/segmental
Glyma.05G002600	346	Known	RAF	WGD/segmental
Glyma.05G036600	391	Known	RAF	WGD/segmental
Glyma.05G037200	352	Known	RAF	WGD/segmental
Glyma.05G167600	475	Known	RAF	WGD/segmental
Glyma.05G220200	416	Known	RAF	WGD/segmental
Glyma.05G245300	1016	Known	RAF	WGD/segmental
Glyma.06G097700	448	Known	RAF	WGD/segmental
Glyma.06G177700	352	Known	RAF	WGD/segmental
Glyma.06G182900	342	New	RAF	WGD/segmental
Glyma.06G183500	386	Known	RAF	WGD/segmental
Glyma.06G276900	812	Known	RAF	WGD/segmental
Glyma.07G101600	1026	Known	RAF	WGD/segmental
Glyma.07G197200	498	Known	RAF	WGD/segmental
Glyma.07G228000	421	Known	RAF	WGD/segmental
Glyma.07G239600	770	Known	RAF	WGD/segmental
Glyma.07G264700	381	Known	RAF	WGD/segmental
Glyma.08G026500	416	Known	RAF	WGD/segmental
Glyma.08G052700	1017	Known	RAF	WGD/segmental
Glyma.08G126000	475	Known	RAF	WGD/segmental
Glyma.08G151400	328	Known	RAF	WGD/segmental
Glyma.08G165800	1245	Known	RAF	WGD/segmental
Glyma.08G165900	1253	Known	RAF	WGD/segmental
Glyma.08G237200	508	New	RAF	Dispersed
Glyma.08G356500	1290	Known	RAF	WGD/segmental
Glyma.09G009100	377	Known	RAF	WGD/segmental
Glyma.09G035400	725	Known	RAF	WGD/segmental
Glyma.09G177600	1022	Known	RAF	WGD/segmental
Glyma.09G275900	370	Known	RAF	WGD/segmental
Glyma.10G066000	836	Known	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Glyma.10G104500	381	New	RAF	Dispersed
Glyma.10G159200	930	Known	RAF	WGD/segmental
Glyma.10G192700	1178	Known	RAF	WGD/segmental
Glyma.10G226300	583	Known	RAF	WGD/segmental
Glyma.10G284400	585	Known	RAF	WGD/segmental
Glyma.11G007100	385	Known	RAF	WGD/segmental
Glyma.11G082200	620	Known	RAF	WGD/segmental
Glyma.12G128700	815	Known	RAF	WGD/segmental
Glyma.12G211000	810	Known	RAF	WGD/segmental
Glyma.13G094500	1110	Known	RAF	WGD/segmental
Glyma.13G151100	836	Known	RAF	WGD/segmental
Glyma.13G169300	366	New	RAF	WGD/segmental
Glyma.13G179000	494	Known	RAF	WGD/segmental
Glyma.13G223400	455	Known	RAF	WGD/segmental
Glyma.13G238400	463	New	RAF	WGD/segmental
Glyma.13G290400	810	Known	RAF	WGD/segmental
Glyma.14G026700	453	Known	RAF	WGD/segmental
Glyma.14G094400	924	Known	RAF	WGD/segmental
Glyma.14G182700	952	Known	RAF	WGD/segmental
Glyma.15G074900	462	Known	RAF	WGD/segmental
Glyma.15G088500	456	Known	RAF	WGD/segmental
Glyma.15G113500	378	Known	RAF	WGD/segmental
Glyma.15G202000	1411	Known	RAF	WGD/segmental
Glyma.15G213400	1222	Known	RAF	WGD/segmental
Glyma.15G261100	1252	Known	RAF	WGD/segmental
Glyma.15G261200	1243	Known	RAF	Tandem
Glyma.15G270700	328	Known	RAF	WGD/segmental
Glyma.15G271100	328	Known	RAF	Proximal
Glyma.16G069200	349	Known	RAF	WGD/segmental
Glyma.17G009300	381	Known	RAF	WGD/segmental
Glyma.17G033700	771	Known	RAF	WGD/segmental
Glyma.17G065700	1096	Known	RAF	WGD/segmental
Glyma.17G090000	359	Known	RAF	WGD/segmental
Glyma.17G090600	392	Known	RAF	WGD/segmental
Glyma.17G105100	1388	Known	RAF	WGD/segmental
Glyma.17G229100	933	Known	RAF	WGD/segmental
Glyma.18G174200	1292	Known	RAF	WGD/segmental
Glyma.19G002800	311	Known	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Glyma.19G009500	367	Known	RAF	WGD/segmental
Glyma.19G056300	348	Known	RAF	WGD/segmental
Glyma.19G191600	808	Known	RAF	WGD/segmental
Glyma.20G031300	423	Known	RAF	WGD/segmental
Glyma.20G105300	583	Known	RAF	WGD/segmental
Glyma.20G149000	381	Known	RAF	WGD/segmental
Glyma.20G165800	557	Known	RAF	WGD/segmental
Glyma.20G197800	1169	Known	RAF	WGD/segmental
Glyma.20G229500	972	Known	RAF	WGD/segmental
Gohir.1Z073500	937	New	RAF	N/A
Gohir.1Z092600	321	New	RAF	N/A
Gohir.A01G001900	570	New	RAF	WGD/segmental
Gohir.A01G099800	1220	New	RAF	WGD/segmental
Gohir.A01G151500	1038	New	RAF	Dispersed
Gohir.A01G171300	640	New	RAF	WGD/segmental
Gohir.A02G003900	353	New	RAF	WGD/segmental
Gohir.A02G007000	474	New	RAF	WGD/segmental
Gohir.A02G024800	1086	New	RAF	WGD/segmental
Gohir.A02G033700	383	New	RAF	WGD/segmental
Gohir.A02G054800	740	New	RAF	WGD/segmental
Gohir.A02G064900	552	New	RAF	WGD/segmental
Gohir.A02G175000	571	New	RAF	WGD/segmental
Gohir.A03G017700	747	New	RAF	WGD/segmental
Gohir.A03G057000	274	New	RAF	Dispersed
Gohir.A03G096700	346	New	RAF	WGD/segmental
Gohir.A03G102400	429	New	RAF	WGD/segmental
Gohir.A04G066300	851	New	RAF	WGD/segmental
Gohir.A04G132100	581	New	RAF	WGD/segmental
Gohir.A05G005800	1081	New	RAF	WGD/segmental
Gohir.A05G011400	1401	New	RAF	WGD/segmental
Gohir.A05G042400	935	New	RAF	WGD/segmental
Gohir.A05G066500	419	New	RAF	WGD/segmental
Gohir.A05G161300	500	New	RAF	WGD/segmental
Gohir.A05G182300	546	New	RAF	WGD/segmental
Gohir.A05G218900	949	New	RAF	WGD/segmental
Gohir.A05G242600	619	New	RAF	WGD/segmental
Gohir.A05G355100	352	New	RAF	WGD/segmental
Gohir.A05G359500	374	New	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Gohir.A05G374000	782	New	RAF	WGD/segmental
Gohir.A05G377800	458	New	RAF	WGD/segmental
Gohir.A06G031600	915	New	RAF	WGD/segmental
Gohir.A06G050100	1361	New	RAF	WGD/segmental
Gohir.A06G089200	920	New	RAF	WGD/segmental
Gohir.A06G123300	457	New	RAF	Dispersed
Gohir.A06G152500	371	New	RAF	WGD/segmental
Gohir.A07G001100	415	New	RAF	WGD/segmental
Gohir.A07G155400	1230	New	RAF	WGD/segmental
Gohir.A07G219500	379	New	RAF	WGD/segmental
Gohir.A07G223900	354	New	RAF	WGD/segmental
Gohir.A08G065000	380	New	RAF	WGD/segmental
Gohir.A08G086100	1034	New	RAF	Dispersed
Gohir.A08G193900	1277	New	RAF	WGD/segmental
Gohir.A08G214000	1012	New	RAF	WGD/segmental
Gohir.A09G029800	351	New	RAF	WGD/segmental
Gohir.A09G066200	765	New	RAF	WGD/segmental
Gohir.A09G135600	846	New	RAF	WGD/segmental
Gohir.A09G218000	391	New	RAF	WGD/segmental
Gohir.A09G234600	777	New	RAF	WGD/segmental
Gohir.A10G009900	374	New	RAF	WGD/segmental
Gohir.A10G026400	486	New	RAF	WGD/segmental
Gohir.A10G033200	453	New	RAF	WGD/segmental
Gohir.A10G146900	472	New	RAF	WGD/segmental
Gohir.A10G172600	775	New	RAF	WGD/segmental
Gohir.A11G067500	383	New	RAF	WGD/segmental
Gohir.A11G171500	576	New	RAF	WGD/segmental
Gohir.A11G186200	386	New	RAF	WGD/segmental
Gohir.A11G256500	998	New	RAF	WGD/segmental
Gohir.A12G003500	390	New	RAF	WGD/segmental
Gohir.A12G047900	1311	New	RAF	WGD/segmental
Gohir.A12G109400	414	New	RAF	WGD/segmental
Gohir.A12G163500	437	New	RAF	WGD/segmental
Gohir.A12G268600	1005	New	RAF	WGD/segmental
Gohir.A12G273500	405	New	RAF	WGD/segmental
Gohir.A13G085100	422	New	RAF	Dispersed
Gohir.D01G001500	552	New	RAF	WGD/segmental
Gohir.D01G084300	1215	New	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Gohir.D01G129100	859	New	RAF	WGD/segmental
Gohir.D01G162300	575	New	RAF	WGD/segmental
Gohir.D02G004300	353	New	RAF	WGD/segmental
Gohir.D02G006600	475	New	RAF	WGD/segmental
Gohir.D02G032000	1096	New	RAF	WGD/segmental
Gohir.D02G040300	421	New	RAF	WGD/segmental
Gohir.D02G059700	737	New	RAF	WGD/segmental
Gohir.D02G059800	725	New	RAF	Tandem
Gohir.D02G070800	552	New	RAF	WGD/segmental
Gohir.D02G117000	346	New	RAF	WGD/segmental
Gohir.D02G127300	430	New	RAF	WGD/segmental
Gohir.D03G004800	571	New	RAF	WGD/segmental
Gohir.D03G153900	747	New	RAF	WGD/segmental
Gohir.D04G037900	458	New	RAF	WGD/segmental
Gohir.D04G041500	782	New	RAF	WGD/segmental
Gohir.D04G054600	374	New	RAF	WGD/segmental
Gohir.D04G058100	352	New	RAF	WGD/segmental
Gohir.D04G104400	851	New	RAF	WGD/segmental
Gohir.D04G168700	581	New	RAF	WGD/segmental
Gohir.D05G006500	1074	New	RAF	WGD/segmental
Gohir.D05G012100	1403	New	RAF	WGD/segmental
Gohir.D05G044100	931	New	RAF	WGD/segmental
Gohir.D05G069000	419	New	RAF	WGD/segmental
Gohir.D05G164100	500	New	RAF	WGD/segmental
Gohir.D05G185300	546	New	RAF	WGD/segmental
Gohir.D05G221600	948	New	RAF	WGD/segmental
Gohir.D05G244000	557	New	RAF	WGD/segmental
Gohir.D05G289200	746	New	RAF	WGD/segmental
Gohir.D06G030800	915	New	RAF	WGD/segmental
Gohir.D06G049400	1032	New	RAF	WGD/segmental
Gohir.D06G087700	903	New	RAF	WGD/segmental
Gohir.D06G140100	457	New	RAF	WGD/segmental
Gohir.D06G158800	416	New	RAF	WGD/segmental
Gohir.D06G196600	723	New	RAF	WGD/segmental
Gohir.D07G001500	415	New	RAF	WGD/segmental
Gohir.D07G022100	937	New	RAF	WGD/segmental
Gohir.D07G161700	1286	New	RAF	WGD/segmental
Gohir.D07G226400	379	New	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Gohir.D07G231000	354	New	RAF	WGD/segmental
Gohir.D08G073500	381	New	RAF	WGD/segmental
Gohir.D08G118400	1038	New	RAF	Dispersed
Gohir.D08G211700	1279	New	RAF	WGD/segmental
Gohir.D08G231000	1013	New	RAF	WGD/segmental
Gohir.D09G029500	368	New	RAF	WGD/segmental
Gohir.D09G065400	765	New	RAF	WGD/segmental
Gohir.D09G137700	852	New	RAF	WGD/segmental
Gohir.D09G208700	470	New	RAF	WGD/segmental
Gohir.D09G210000	355	New	RAF	WGD/segmental
Gohir.D09G213000	470	New	RAF	WGD/segmental
Gohir.D09G214400	355	New	RAF	WGD/segmental
Gohir.D09G220500	391	New	RAF	WGD/segmental
Gohir.D09G235400	776	New	RAF	WGD/segmental
Gohir.D10G009600	374	New	RAF	WGD/segmental
Gohir.D10G027000	486	New	RAF	WGD/segmental
Gohir.D10G033900	453	New	RAF	WGD/segmental
Gohir.D10G119200	472	New	RAF	WGD/segmental
Gohir.D10G150400	391	New	RAF	WGD/segmental
Gohir.D10G179100	775	New	RAF	WGD/segmental
Gohir.D11G071400	377	New	RAF	WGD/segmental
Gohir.D11G130100	391	New	RAF	WGD/segmental
Gohir.D11G192800	386	New	RAF	WGD/segmental
Gohir.D11G266300	1000	New	RAF	WGD/segmental
Gohir.D12G003200	390	New	RAF	WGD/segmental
Gohir.D12G046700	1311	New	RAF	WGD/segmental
Gohir.D12G112400	414	New	RAF	WGD/segmental
Gohir.D12G166500	427	New	RAF	WGD/segmental
Gohir.D12G269000	1006	New	RAF	WGD/segmental
Gohir.D12G274200	399	New	RAF	WGD/segmental
Gohir.D13G077900	422	New	RAF	WGD/segmental
Gorai.001G001200	415	Known	RAF	Dispersed
Gorai.001G022800	937	Known	RAF	WGD/segmental
Gorai.001G185100	1320	New	RAF	WGD/segmental
Gorai.001G261300	381	Known	RAF	WGD/segmental
Gorai.001G266100	354	Known	RAF	WGD/segmental
Gorai.002G002100	552	Known	RAF	WGD/segmental
Gorai.002G110000	1250	New	RAF	Dispersed

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Gorai.002G159900	859	Known	RAF	WGD/segmental
Gorai.002G201200	575	Known	RAF	WGD/segmental
Gorai.003G005100	571	Known	RAF	WGD/segmental
Gorai.003G162100	757	New	RAF	Dispersed
Gorai.004G077900	381	Known	RAF	WGD/segmental
Gorai.004G101000	1137	Known	RAF	WGD/segmental
Gorai.004G225000	1277	New	RAF	WGD/segmental
Gorai.004G245400	1013	New	RAF	WGD/segmental
Gorai.005G004900	353	Known	RAF	WGD/segmental
Gorai.005G007700	475	Known	RAF	WGD/segmental
Gorai.005G035500	1096	New	RAF	Dispersed
Gorai.005G044700	383	Known	RAF	WGD/segmental
Gorai.005G067300	740	New	RAF	WGD/segmental
Gorai.005G067400	739	Known	RAF	Tandem
Gorai.005G080000	553	New	RAF	WGD/segmental
Gorai.005G136700	346	New	RAF	Dispersed
Gorai.005G142900	430	Known	RAF	WGD/segmental
Gorai.006G034500	351	Known	RAF	WGD/segmental
Gorai.006G056600	1038	Known	RAF	WGD/segmental
Gorai.006G080100	767	Known	RAF	WGD/segmental
Gorai.006G158200	851	Known	RAF	WGD/segmental
Gorai.006G233100	509	Known	RAF	WGD/segmental
Gorai.006G234800	355	Known	RAF	WGD/segmental
Gorai.006G241500	391	Known	RAF	WGD/segmental
Gorai.006G258400	804	Known	RAF	WGD/segmental
Gorai.007G076700	377	Known	RAF	WGD/segmental
Gorai.007G139200	391	Known	RAF	WGD/segmental
Gorai.007G190200	576	Known	RAF	WGD/segmental
Gorai.007G205800	398	New	RAF	Dispersed
Gorai.007G286200	1000	New	RAF	WGD/segmental
Gorai.008G003000	390	Known	RAF	WGD/segmental
Gorai.008G049200	1315	New	RAF	Dispersed
Gorai.008G121800	414	New	RAF	Dispersed
Gorai.008G179300	427	Known	RAF	WGD/segmental
Gorai.008G289300	1007	New	RAF	WGD/segmental
Gorai.008G294700	399	Known	RAF	Dispersed
Gorai.009G006600	1107	Known	RAF	WGD/segmental
Gorai.009G012400	1403	New	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Gorai.009G045200	923	Known	RAF	WGD/segmental
Gorai.009G069900	419	New	RAF	Dispersed
Gorai.009G169300	500	New	RAF	WGD/segmental
Gorai.009G190600	546	Known	RAF	WGD/segmental
Gorai.009G228800	948	New	RAF	WGD/segmental
Gorai.009G253200	556	New	RAF	WGD/segmental
Gorai.009G304300	786	New	RAF	Dispersed
Gorai.009G408200	851	Known	RAF	WGD/segmental
Gorai.010G033900	915	Known	RAF	WGD/segmental
Gorai.010G056300	1360	Known	RAF	WGD/segmental
Gorai.010G101300	919	New	RAF	WGD/segmental
Gorai.010G163500	457	New	RAF	WGD/segmental
Gorai.010G185600	371	Known	RAF	WGD/segmental
Gorai.011G010500	374	Known	RAF	WGD/segmental
Gorai.011G028300	486	New	RAF	WGD/segmental
Gorai.011G035600	453	New	RAF	WGD/segmental
Gorai.011G131000	472	New	RAF	WGD/segmental
Gorai.011G167400	391	Known	RAF	WGD/segmental
Gorai.011G200300	775	Known	RAF	WGD/segmental
Gorai.012G042800	458	New	RAF	Dispersed
Gorai.012G046800	782	Known	RAF	Dispersed
Gorai.012G061300	374	Known	RAF	WGD/segmental
Gorai.012G064800	353	Known	RAF	WGD/segmental
Gorai.012G157000	591	Known	RAF	WGD/segmental
Gorai.013G089100	422	Known	RAF	WGD/segmental
GRMZM2G002531	534	New	RAF	WGD/segmental
GRMZM2G007854	787	Known	RAF	Dispersed
GRMZM2G011070	1221	New	RAF	N/A
GRMZM2G014618	442	Known	RAF	WGD/segmental
GRMZM2G018280	404	Known	RAF	WGD/segmental
GRMZM2G019434	370	Known	RAF	Dispersed
GRMZM2G028604	396	Known	RAF	WGD/segmental
GRMZM2G028709	580	New	RAF	Dispersed
GRMZM2G038982	903	Known	RAF	WGD/segmental
GRMZM2G039106	1139	Known	RAF	WGD/segmental
GRMZM2G045366	471	Known	RAF	Dispersed
GRMZM2G048243	1071	Known	RAF	Dispersed
GRMZM2G052658	1104	Known	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
GRMZM2G055334	574	Known	RAF	WGD/segmental
GRMZM2G059671	800	Known	RAF	Dispersed
GRMZM2G063069	377	Known	RAF	WGD/segmental
GRMZM2G063684	382	Known	RAF	WGD/segmental
GRMZM2G072584	412	New	RAF	Dispersed
GRMZM2G080499	792	Known	RAF	WGD/segmental
GRMZM2G088299	382	Known	RAF	WGD/segmental
GRMZM2G097878	561	Known	RAF	WGD/segmental
GRMZM2G098187	762	Known	RAF	Dispersed
GRMZM2G102088	415	Known	RAF	WGD/segmental
GRMZM2G104283	602	Known	RAF	Dispersed
GRMZM2G104658	562	New	RAF	Dispersed
GRMZM2G110572	752	Known	RAF	WGD/segmental
GRMZM2G111269	378	Known	RAF	Dispersed
GRMZM2G114093	598	Known	RAF	Dispersed
GRMZM2G127632	1032	New	RAF	Dispersed
GRMZM2G131629	416	Known	RAF	WGD/segmental
GRMZM2G140537	825	Known	RAF	WGD/segmental
GRMZM2G140612	423	Known	RAF	WGD/segmental
GRMZM2G152889	525	Known	RAF	WGD/segmental
GRMZM2G156013	415	Known	RAF	WGD/segmental
GRMZM2G159034	440	Known	RAF	Dispersed
GRMZM2G160922	531	Known	RAF	Dispersed
GRMZM2G163141	791	Known	RAF	WGD/segmental
GRMZM2G164242	569	Known	RAF	Dispersed
GRMZM2G165231	353	Known	RAF	Dispersed
GRMZM2G171677	368	Known	RAF	Dispersed
GRMZM2G175563	892	Known	RAF	WGD/segmental
GRMZM2G326472	1114	Known	RAF	Dispersed
GRMZM2G413069	869	Known	RAF	Dispersed
GRMZM2G448213	675	Known	RAF	Dispersed
GRMZM2G459854	593	Known	RAF	WGD/segmental
GRMZM2G465833	529	Known	RAF	WGD/segmental
GRMZM2G474546	593	Known	RAF	WGD/segmental
GRMZM2G481005	1265	Known	RAF	Dispersed
GRMZM5G814851	594	Known	RAF	WGD/segmental
Solyc01g010950.2	430	Known	RAF	Dispersed
Solyc01g059860.2	760	Known	RAF	Dispersed

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Solyc01g097980.2	982	Known	RAF	Dispersed
Solyc01g111880.2	563	Known	RAF	WGD/segmental
Solyc02g031860.2	1221	New	RAF	Proximal
Solyc02g031910.2	1242	New	RAF	Proximal
Solyc02g071740.2	461	Known	RAF	WGD/segmental
Solyc02g076780.2	741	Known	RAF	Dispersed
Solyc02g078140.2	504	Known	RAF	WGD/segmental
Solyc02g083290.2	562	New	RAF	WGD/segmental
Solyc02g085620.2	371	New	RAF	Dispersed
Solyc02g093410.2	353	Known	RAF	Dispersed
Solyc03g005920.2	462	New	RAF	WGD/segmental
Solyc03g006400.2	480	Known	RAF	WGD/segmental
Solyc03g114210.2	391	New	RAF	WGD/segmental
Solyc03g114310.2	351	Known	RAF	Dispersed
Solyc03g119140.2	1031	Known	RAF	Dispersed
Solyc03g121780.1	311	Known	RAF	Dispersed
Solyc04g014690.2	377	Known	RAF	WGD/segmental
Solyc04g071120.2	541	New	RAF	Dispersed
Solyc04g076480.2	958	Known	RAF	Dispersed
Solyc05g013070.2	436	New	RAF	Dispersed
Solyc06g068980.2	989	Known	RAF	Dispersed
Solyc06g071410.2	394	Known	RAF	WGD/segmental
Solyc07g006760.2	1083	Known	RAF	Dispersed
Solyc07g007140.2	1415	Known	RAF	Dispersed
Solyc07g042680.2	412	Known	RAF	WGD/segmental
Solyc07g042890.2	412	Known	RAF	Dispersed
Solyc07g055130.2	813	Known	RAF	Dispersed
Solyc07g055870.2	459	Known	RAF	Dispersed
Solyc08g007910.2	723	Known	RAF	WGD/segmental
Solyc08g014450.2	472	New	RAF	Dispersed
Solyc08g065250.2	746	New	RAF	Dispersed
Solyc08g080460.1	756	Known	RAF	WGD/segmental
Solyc09g009090.2	837	Known	RAF	WGD/segmental
Solyc09g018060.2	310	Known	RAF	Dispersed
Solyc09g082470.2	395	New	RAF	Dispersed
Solyc09g091460.2	1235	New	RAF	Dispersed
Solyc10g017490.1	439	Known	RAF	Dispersed
Solyc10g055720.1	563	Known	RAF	WGD/segmental

Table A.1 (continued)

Gene ID	Length	Identity	Subfamily	Duplicate type
Solyc10g083610.1	829	Known	RAF	WGD/segmental
Solyc10g085570.1	793	Known	RAF	WGD/segmental
Solyc11g012050.1	374	Known	RAF	WGD/segmental
Solyc12g009340.1	400	Known	RAF	WGD/segmental
Solyc12g013980.1	391	Known	RAF	Dispersed
Solyc12g019410.1	469	Known	RAF	Dispersed
Solyc12g062280.1	362	New	RAF	Dispersed
Solyc12g094410.1	569	New	RAF	Dispersed
Solyc12g099250.1	766	Known	RAF	Dispersed
Zosma107g00710	592	New	RAF	N/A
Zosma117g00130	391	New	RAF	N/A
Zosma146g00510	520	New	RAF	N/A
Zosma155g00360	371	New	RAF	N/A
Zosma160g00390	447	New	RAF	N/A
Zosma173g00130	573	New	RAF	N/A
Zosma199g00230	359	New	RAF	N/A
Zosma19g00030	1400	New	RAF	N/A
Zosma19g00420	405	New	RAF	N/A
Zosma1g01920	1299	New	RAF	N/A
Zosma216g00040	745	New	RAF	N/A
Zosma261g00180	368	New	RAF	N/A
Zosma289g00100	1019	New	RAF	N/A
Zosma31g00240	469	New	RAF	N/A
Zosma3g01460	922	New	RAF	N/A
Zosma50g00950	1042	New	RAF	N/A
Zosma58g00140	992	New	RAF	N/A
Zosma59g00800	521	New	RAF	N/A
Zosma62g00330	814	New	RAF	N/A
Zosma63g00330	353	New	RAF	N/A
Zosma63g00600	688	New	RAF	N/A
Zosma64g00150	557	New	RAF	N/A
Zosma69g00340	1219	New	RAF	N/A
Zosma91g00270	436	New	RAF	N/A
Zosma9g00610	366	New	RAF	N/A

Table of all identified MAP3Ks in seven plant species. The columns from left to right indicate gene identifiers, length of the longest transcript variant used for all analyses, novelty of the protein compared to previous MAP3K examinations, subfamily localization, and gene origin classification generated by MCScanX.

APPENDIX B

SUPPLEMENTS FOR CHAPTER III

Table B.1 Table of primers used for VIGS and qRT-PCR for *GhAct7*

Type	Name	Sequence	Amplicon Size (bp)	Insert Size (bp)
VIGS	GhAct7_F	CGTTAACCTCACATCGTGCCAATCTATG	309	294
VIGS	GhAct7_R	GGACTAGTGGCAACGGAATCTCTCAGCT		
qRT-PCR	RT_GhAct7_F	CCTCCGTCTAGACCTTGCTG	158	
qRT-PCR	RT_GhAct7_R	CCAGCTCCTGCTCATAGTCC		

Primers used to amplify VIGS silencing fragment for *GhAct7* and the primers used to check *GhAct7* expression using qRT-PCR. Expected PCR amplicon size and VIGS insert size is indicated beside each set of primers. The "type" column indicates the function of each primer.

Table B.2 Table of primers used for VIGS and qRT-PCR for *GhILK1.1*

Type	Name	Sequence	Amplicon Size (bp)	Insert Size (bp)
VIGS	GhILK1.1_F	GCGTTAACCTCAAATAGGGCGTTAAAACG	190	174
VIGS	GhILK1.1_R	CCACTAGTATGGTGGAGTCCAAATTCTCC		
qRT-PCR	RT_GhILK1.1_F	CCAGGAAGACGCCAATGACAG	81	
qRT-PCR	RT_GhILK1.1_R	TCCGAACTTGGAGCTCGACTG		

Primers used to amplify VIGS silencing fragment for *GhILK1.1* and the primers used to check *GhILK1.1* expression using qRT-PCR. Expected PCR amplicon size and VIGS insert size is indicated beside each set of primers. The "type" column indicates the function of each primer.

Table B.3 Table of primers used for VIGS and qRT-PCR for *GhILK1.2*

Type	Name	Sequence	Amplicon Size (bp)	Insert Size (bp)
VIGS	GhILK1.2_F	GGGGTTAACATGGAGAACTTAGCATCGCAG	215/218	198/201
VIGS	GhILK1.2_R	GGACTAGTCGGCACCGACATCTCATCTT		
qRT-PCR	RT_GhILK1.2_F	ATCCTCAAGAGGTGCCTGAG	70	
qRT-PCR	RT_GhILK1.2_R	GTTATACCGTCAGACTTCCGAA		

Primers used to amplify VIGS silencing fragment for *GhILK1.2* and the primers used to check *GhILK1.2* expression using qRT-PCR. Expected PCR amplicon size and VIGS insert size is indicated beside each set of primers. The "type" column indicates the function of each primer.

Table B.4 Table of primers for VIGS and qRT-PCR for *GhILK1.3*

Type	Name	Sequence	Amplicon Size (bp)	Insert Size (bp)
VIGS	GhILK1.3_F	GGGGTTAACTCAGTTAAACCGGGGAATCTC	212	195
VIGS	GhILK1.3_R	GGACTAGTTCAAGGTTTTCCGGCACC		
qRT-PCR	RT_GhILK1.3_F	CAATGGCAGTCACAAATCCTC	80	
qRT-PCR	RT_GhILK1.3_R	ACCATCAGACTTCCGAACTTG		

Primers used to amplify VIGS silencing fragment for *GhILK1.3* and the primers used to check *GhILK1.3* expression using qRT-PCR. Expected PCR amplicon size and VIGS insert size is indicated beside each set of primers. The "type" column indicates the function of each primer.

APPENDIX C

EXAMINATION OF SWEET POTATO PEPTIDE FRAGMENTS

Background

Ipomea batatas is a critically important global commercial crop whose genome has – until recently - escaped serviceable assembly and annotation due to its highly polymorphic allo-autohexaploid genome (84). A reference sweet potato proteome is also presently lacking. In order to evaluate the performance of two methods for protein extraction from sweet potato leaves and tuberous roots, mass spectrometry-identified peptide fragments were BLASTed against the Ipomoea taxon on NCBI. The resulting top hits were subjected to gene ontology analyses in order to explore the composition of leaf and root proteomes and evaluate the efficacy of two distinct methods of protein extraction and solubilization (85).

Methods

Peptide spectra from mass-spectrometry identified proteins were previously matched against the NCBI Ipomoea taxon (taxid: 4119) protein dataset using MASCOT v2.4. Protein coding sequences were retrieved for all 4,321 unique hits using NCBI. The PANTHER HMM Scoring tool was used to score protein sequences against the entire PANTHER HMM library which was last updated 2/8/18 (86). Top HMM hits below an e-value cutoff of 0.001 were kept for further analysis resulting in 4286/4321 (99.2%) sequences being mapped to a PANTHER family. PANTHER Generic Mapping files were generated for individual datasets in R for GO term enrichment. Protein classes were retrieved using the PANTHER.db (v1.0.4) package. GOSLIM terms were identified using the PANTHER GOslim OBO file available from "http://data.pantherdb.org/PANTHER13.1/ontology/PANTHERGOslim.obo" using the onthologyIndex (v2.0) package.

Results

Unique protein identification

4,321 unique proteins were identified using top BLAST hits of the sweet potato peptide fragments. Method 1 (M1; phenol-based protein extraction method) identified 2,681 and 2,641 unique proteins from leaf and root tissue respectively. Method 2 (M2; polyethylene glycol 4000 fractionation-based method) identified 1,589 and 1,368 unique proteins from leaf and root tissue respectively. Both methods were tested on 4 technical replicates within both tissue types and resulted in the identification of similar numbers of proteins, indicating the high reproducibility of the extraction workflow. Independent of method, 3,142 leaf and 2,925 root proteins were identified, representing 4,321 unique proteins – the largest dataset of *I. batata* proteins to date.

An examination of differences between the two methods revealed uniquely extracted proteins in both leaf (1,554 unique proteins identified using M1 but missed by M2; 461 unique proteins identified by M2 and missed by M1) and root (1,558 unique proteins identified by M1 and missed by M2; 285 unique proteins identified by M2 and missed by M1). Overall, although leaf tissue yielded a slightly higher number of unique proteins (3,142 hits independent of method in leaf vs 2,925 from root tissue), M1 surpassed M2 in both the total number of proteins extracted (3,839 vs 2,203 in M2) and the number of proteins extracted from both leaf and root tissue.

Mapping PANTHER classes within extraction methods

PANTHER protein classes were mapped to identified peptides to identify protein classes with significant overrepresentation and differential accumulation across method or tissue type. Comparing M1 with M2 preparations, we found marked differences

regarding both the identity of protein class and number of members per class (Figure C.1). Overall, M1 greatly outperformed M2 in the contribution to members in individual protein classes with 121 classes having more members identified using M1 (average 275 proteins/class), compared to 20 classes receiving more members from M2 (average 170 proteins/ class). Proteins belonging to 152 unique protein classes were identified across the two methods, with 11 classes common between M1 and M2. Cytoskeletal proteins, transferases, transfer/carriers, G-proteins and signaling molecules were among the protein classes with the highest representation in both M1 and M2 preparations. DNA-binding proteins, such as polymerases and centromere-binding factors, and small GTPases were preferentially extracted by M1; on the other hand, extracellular matrix proteins and ribonucleases were present only in M2 preparations.

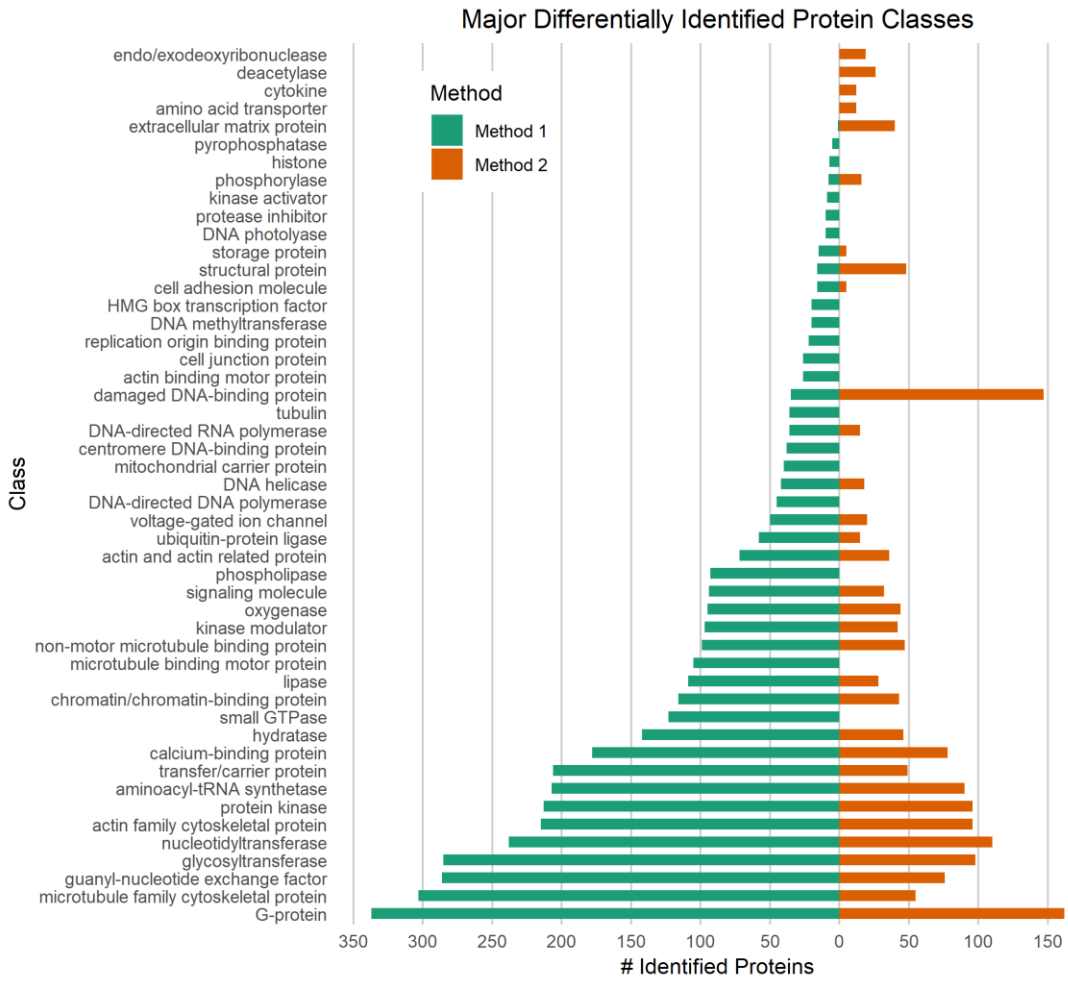


Figure C.1 Comparison of most differentially expressed protein classes between M1 and M2.

GO term descriptions populate the y-axis with the number of proteins identified from either method populating the x-axis. All displayed protein classes have at least a 2-fold difference between M1 and M2.

Mapping PANTHER classes within tissues

Mapping PANTHER classes across tissue types revealed that root tissue allowed for both a larger variety and number of protein classes to be extracted from sweet potato (Figure C.2). Overall, 72 protein classes were preferentially extracted from root tissue, compared to 65 classes from leaf tissue; only 15 protein classes were identified in both tissues. The roots yielded nine unique protein classes, among which enzymes including Ser/Thr receptors, ribonucleases, and polymerases had a good representation. The leaf tissue yielded 13 unique protein classes including G-protein coupled receptors and other non-receptor tyrosine kinases, proteases and protease inhibitors. An enrichment analysis showed significant overrepresentation of GO terms associated with primary metabolic processes and with the localization, intracellular transport, and exocytosis of proteins; in addition, amylases involved in the conversion of starches into simple sugars and SNARE proteins mediating vesicle fusion were 3.5-fold enriched in roots versus leaves preparations. On the other hand, terms associated with protein translation; chromatin remodeling and photosynthetic processes were overrepresented in leaves, indicating actively metabolizing tissue.

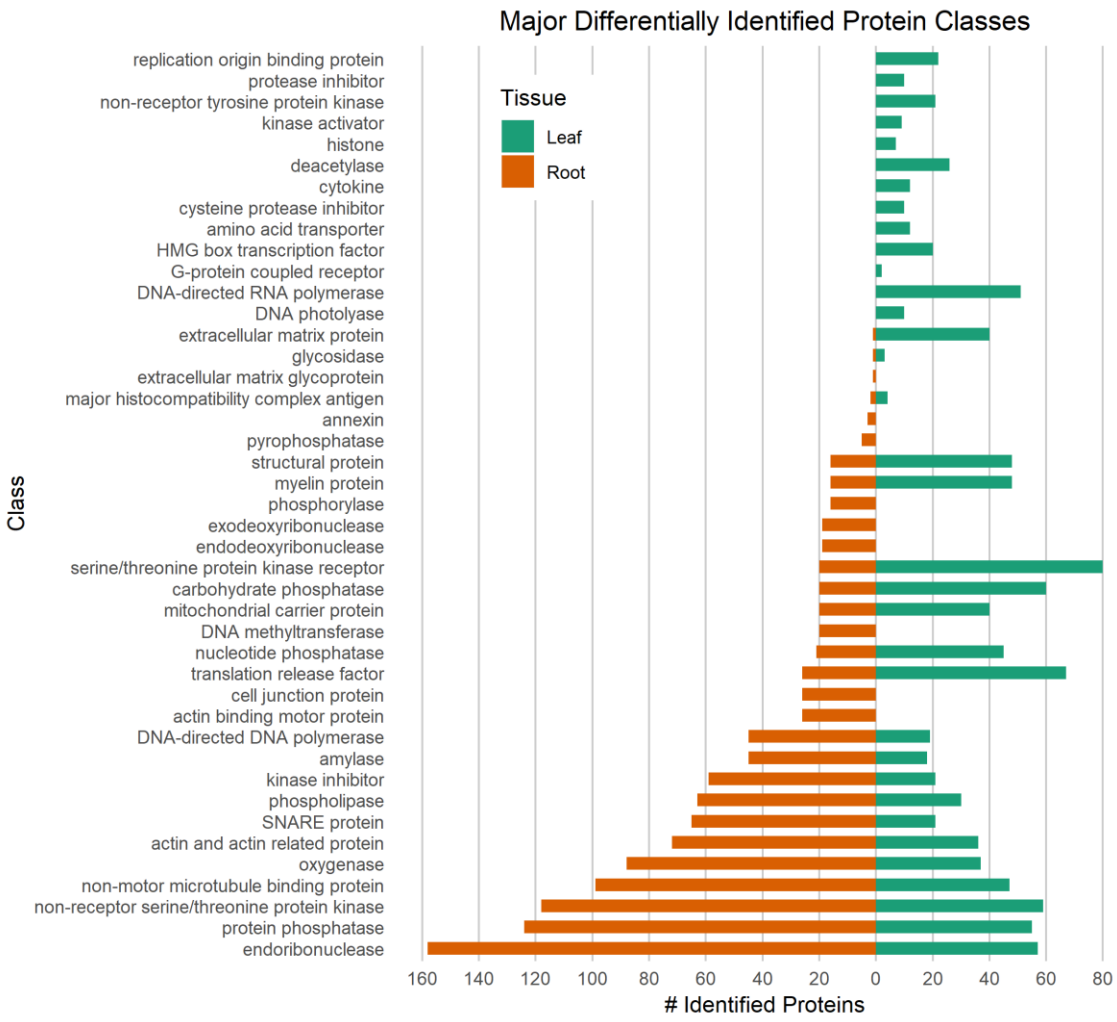


Figure C.2 Comparison of most differentially expressed protein classes between leaf and root tissue.

GO term descriptions populate the y-axis with the number of proteins identified from either tissue populating the x-axis. All displayed protein classes have at least a 2-fold difference between leaf and root tissue.

Examination of Biological Process GO terms across extraction methods

An analysis of GO biological processes (BP) terms across methodologies identified 156 unique terms across M1 and M2 (Figure C.3). M1 greatly outperformed M2 in the identification of proteins within unique BP terms. Of all examined proteins, M1 and M2 identified completely all constituents within 52 and 16 BP terms respectively, with 12 terms equally represented by either method. Further, of the 156 BP terms identified, 154 (98.7%) were populated using hits identified using M1. This is in contrast to the 90 BP terms (57.7%) which were at least halfway populated using hits identified using M2. Further, M1 identified proteins categorized across a wider breadth of BP terms with all 156 terms identified by M1 and only 142 terms identified by M2. M2 failed to identify proteins associated with aspects of growth and development including vitamin transport, phagocytosis, and chromatin remodeling.

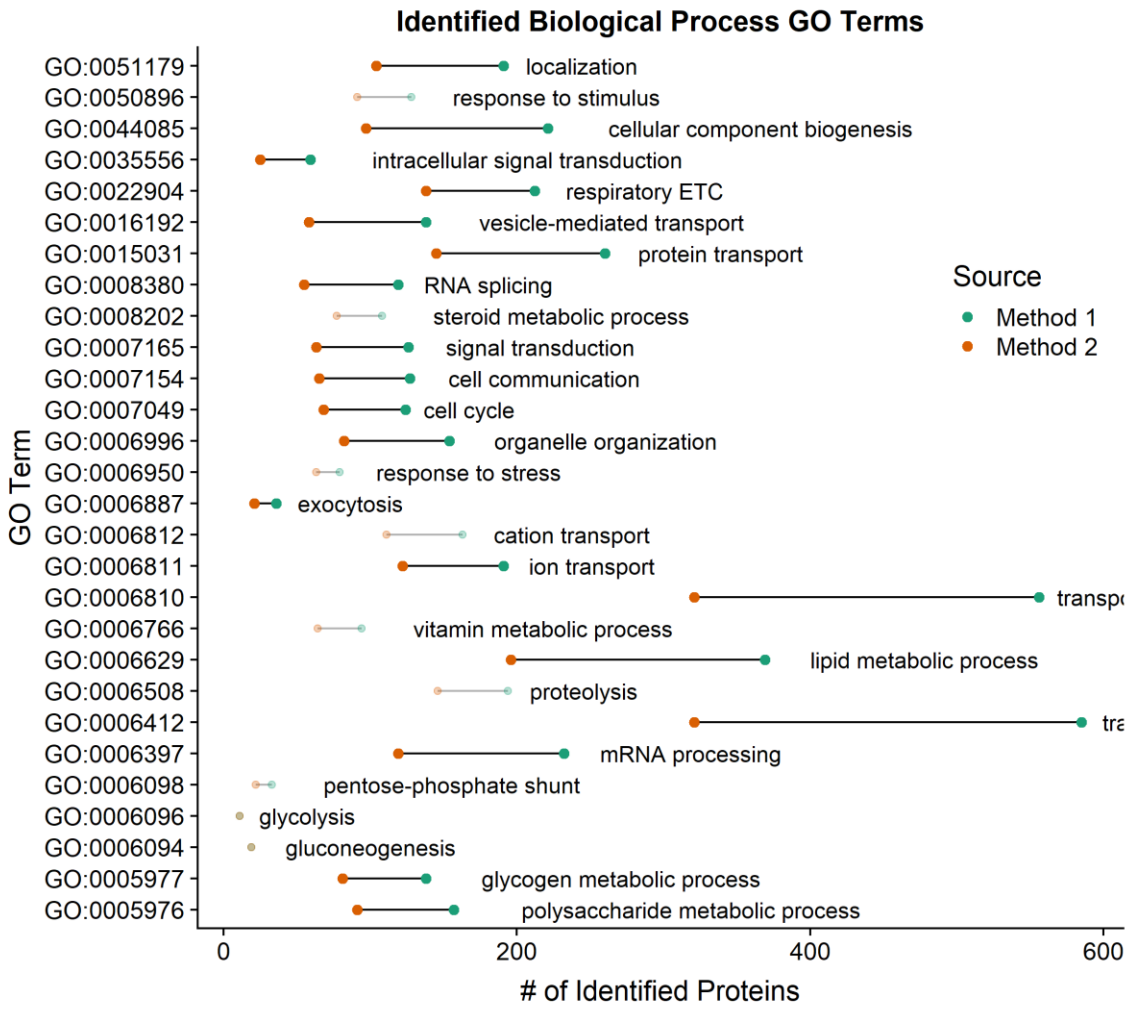


Figure C.3 Differential examination of Biological Process (BP) GO terms identified by M1 and M2.

Dots represent the frequency of uniquely identified proteins from either Method 1 or Method 2. Bars connecting dots represent the difference in the number of proteins extracted by either Method with all highlighted bars showing at least a 50% increase in term population. Plotted terms represent lower hierarchy terms.

Examination of Cellular Component terms across extraction methods

Examining GO cellular component (CC) terms yielded similar observations. Within the 44 identified CC terms, 17 and 1 terms were fully populated using M1 and M2 respectively. All identified CC terms were at least halfway populated using M1, whereas M2 resulted in at least 50% population within only 15 CC terms (34%). M1 exclusively identified proteins associated with the peroxisome (12 proteins), nucleoplasm (15 proteins), and tubulin complexes (6 proteins). Protein transport, mRNA processing and translation, and lipid metabolism were terms showing the largest difference between methods, with more proteins in these GOBP terms extracted by M1. Similarly, M1 extracted more efficiently proteins localized in the membrane, cytosol, nucleus, and ribosomal proteins. While both methods performed equally well for proteins localized within the cell wall , neither M1 nor M2 extracted well proteins associated with the extracellular matrix (Figure C.4).

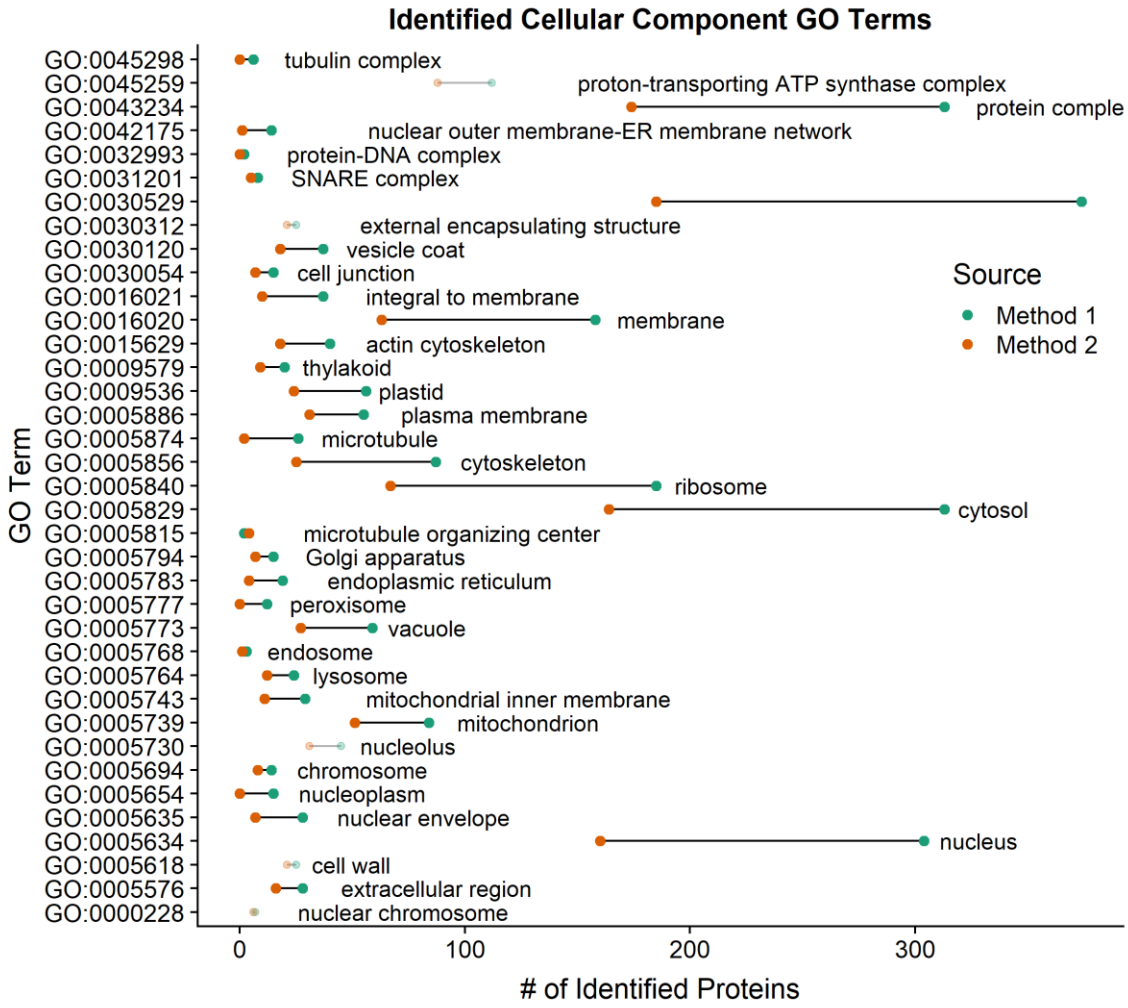


Figure C.4 Differential examination of Cellular Component (CC) GO terms identified by M1 and M2.

Dots represent the frequency of uniquely identified proteins from either Method 1 or Method 2. Bars connecting dots represent the difference in the number of proteins extracted by either Method with all highlighted bars showing at least a 50% increase in term population. Plotted terms represent lower hierarchy terms.

Conclusion

Mass spectrometry-identified peptide fragments were mapped to representative, homologous protein sequences using BLAST. 4,321 unique hits were identified, representing proteins extracted from sweet potato leaf and root tissues using two protein extraction methodologies. Unique hits were mapped to PANTHER families and analyzed to show that Method 1 – a phenol-based protein extraction method – consistently identified a larger number and variety of protein classes compared to Method 2. Gene ontology annotations revealed similar observations regarding Biological Process, Cellular Component, and Molecular Functions between the two methods. Finally, a comparison of tissue types revealed that although a similar number of proteins were extracted from both leaf and root tissues, differences in quantity, mapped protein class, and ontology-annotated function were observed.

APPENDIX D

RNA-SEQ ANALYSIS OF ARABIDOPSIS MUTANT PLANTS WITH ALTERED *ILK1* EXPRESSION

Background

The collection of all RNA expressed by the genome of a cell is termed collectively as its transcriptome. Although not every gene is transcriptionally active at all times (or in all tissues), changes in gene expression can be measured using transcriptome data to assess global transcriptional responses to environmental stresses. Arabidopsis plants with reduced *ILK1* expression were subjected to two distinct treatments – salt stress induced through NaCl and biotic stress induced through the bacterial pathogen-associated molecular pattern flg22. RNA sequencing was utilized to capture transcriptional responses following stress perception at 4 different time points for both stresses. The identification of differentially expressed genes and gene transcripts was presently examined to uncover the role of *ILK1* in mediating salt stress responses and PAMP-triggered immunity.

Methods

Raw reads were initially obtained from Arabidopsis tissues treated with either the flg22 PAMP or NaCl at four distinct time points: 0 hours, 3 hours, 6 hours, and 12 hours after treatment. Original raw read data was filtered using Trimmomatic v0.32. Trimmed reads were aligned to the Araport11 reference genome for Arabidopsis acquired from Phytozome using HISAT2 v2.0.4. FASTQC v0.11.3 was used to assess quality of reads. Transcript assembly and quantitation was performed using Stringtie v1.3.0. Gffcompare v0.10.5 was used to assess quality and precision of transcript assembly. Ballgown v2.12.0 in R v3.5.1 was finally used to perform differential expression analysis.

Results

Results of trimming raw reads

RNA sequencing initially generated around 394 and 421 million 50bp-long pair-ended reads from the NaCl and flg22 treatments. Reads were acquired from three biological replicates at four different time points in both wild type control and *ILK1* knockdown plants for both treatments resulting in a total of 48 samples averaging around 17 million reads per sample. In order to filter lower quality reads from the dataset, raw reads were trimmed using Trimmomatic. First, all base pair calls below a quality score of 28 at the beginning of reads were cut. A sliding window approach examining 4 base pairs at a time - starting at the 5' end of reads - was then implemented to clip individual reads once the average quality within the window fell below 28. Finally, all remaining reads that were not at least 36 base pairs long were removed. This resulted in an average of around 13 million (77%) high-quality, pair-ended reads remaining for each sample.

Mapping trimmed reads and transcript assembly

Trimmed reads were mapped against the Araport11 Arabidopsis reference genome using HISAT2. Around 91% of reads were aligned to a single position within the genome and an average overall alignment rate of 98.6% was reported among all examined samples. Reads were assembled and counted using Stringtie, generating 65,491 transcripts across 26,192 genes. An assessment of the assembled genes and transcripts revealed excellent overlap with the Araport11 reference genome. Although the Araport11 reference genome appears to encode 494 duplicated annotations, of the 27,172 predicted unique genes, Stringtie assembled reads covering 100% of the genes at least partially across all samples and only resulted in the prediction of 623 (2.4% of

predicted genes) novel loci. No reference exons were missed and only a single reference intron was absent, while only 2,646 novel exons (1.2% of all predicted exons) and 2,275 novel introns (1.6% of all predicted introns).

Transcript-level differential expression analysis

Differential expression analysis using Ballgown revealed no significant differential expression for any genes at any time points between *ILK1* knockdown and WT control plants when using multiple testing correction where significant differential expression was set to include any gene with a q-value of at least 0.05 and a $\log_2$ normalized fold change of at least 2 in either the NaCl or the flg22 datasets. Given the reduced transcriptional reprogramming effect associated with *ILK1* knockdown, the change in expression might be better quantified by not correcting significance testing results for changes in expression within the whole genome dataset. Reducing the stringency of the analysis to keep all transcripts with a p-value below 0.05 while maintaining the previous fold change criteria, however, resulted in differentially expressed transcripts resulting from both examined treatments (Figure D.1 and Figure D.2). In total, 170, 439, 297, and 214 assembled transcripts were differentially expressed at 0, 3, 6, and 12 hours respectively following NaCl treatment. Following the flg22 treatment, 235, 160, 241, and 303 transcripts were found to be similarly differentially expressed. 11, 13, 21, and 9 differentially expressed transcripts were shared between the two treatments at 0, 3, 6, and 12 hours respectively (Figure D.3).

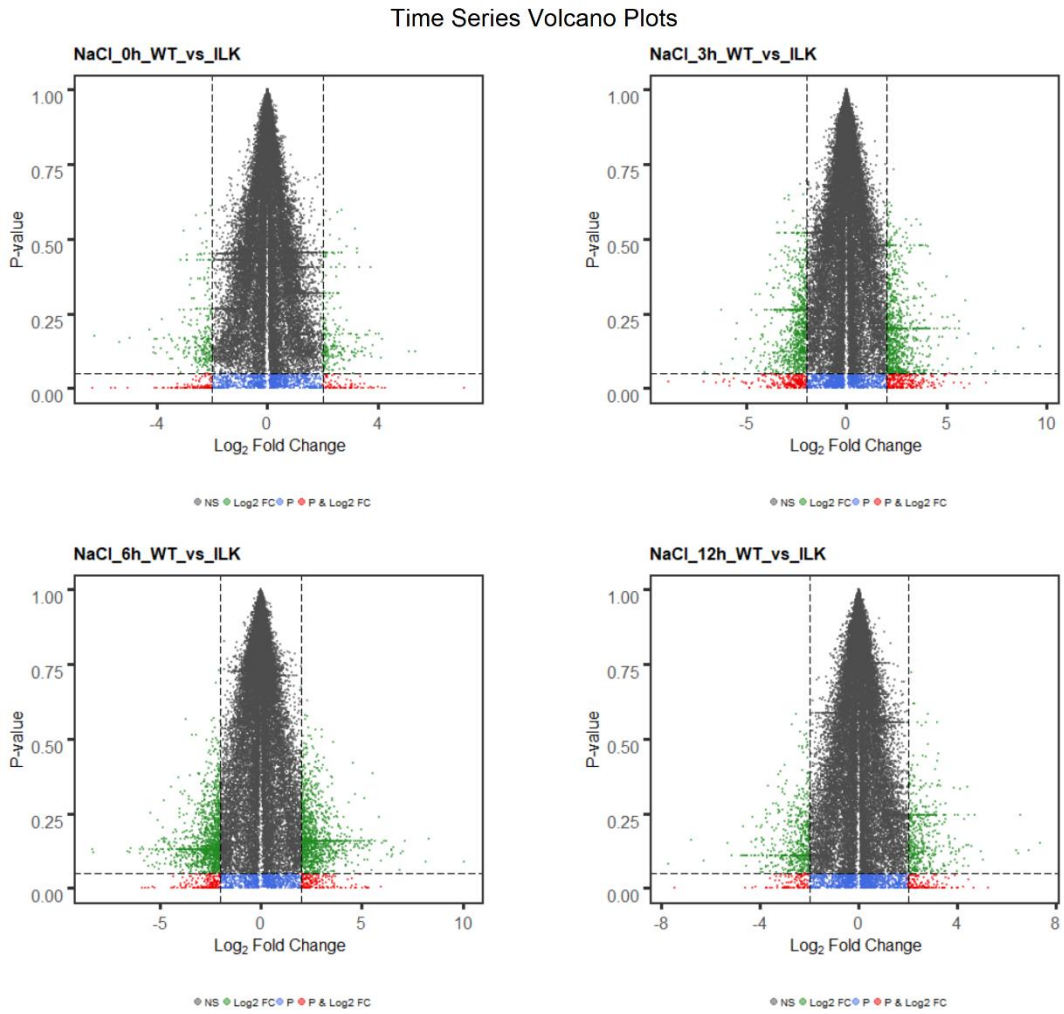


Figure D.1 Transcript-level differences in expression following NaCl treatment between *ILK1* knockdown and wild type control Arabidopsis

For each of the 4 examined time points (0 hour, 3 hours, 6 hours, and 12 hours after treatment), assembled transcripts are plotted in the above volcano plots. The legend below each plot describes the significance of each plotted transcript. Significant differentially expressed transcripts are labelled as red dots and display a $\log_2$(Fold change) of at least 2 and a p-value of at least 0.05.

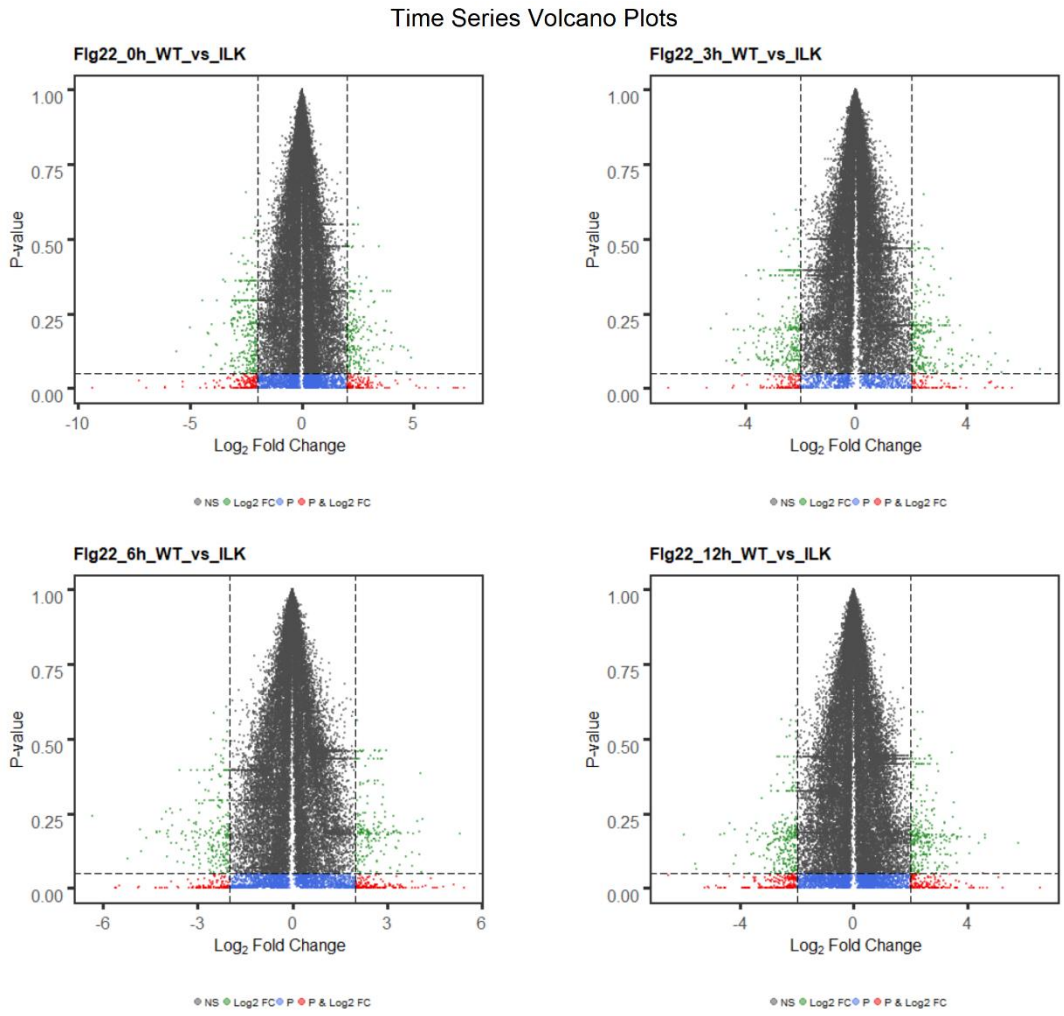


Figure D.2 Transcript-level differences in expression following flg22 perception between *ILK1* knockdown and wild type control Arabidopsis

For each of the 4 examined time points (0 hour, 3 hours, 6 hours, and 12 hours after treatment), assembled transcripts are plotted in the above volcano plots. The legend below each plot describes the significance of each plotted transcript. Significant differentially expressed transcripts are labelled as red dots and display a $\log_2$(Fold change) of at least 2 and a p-value of at least 0.05.

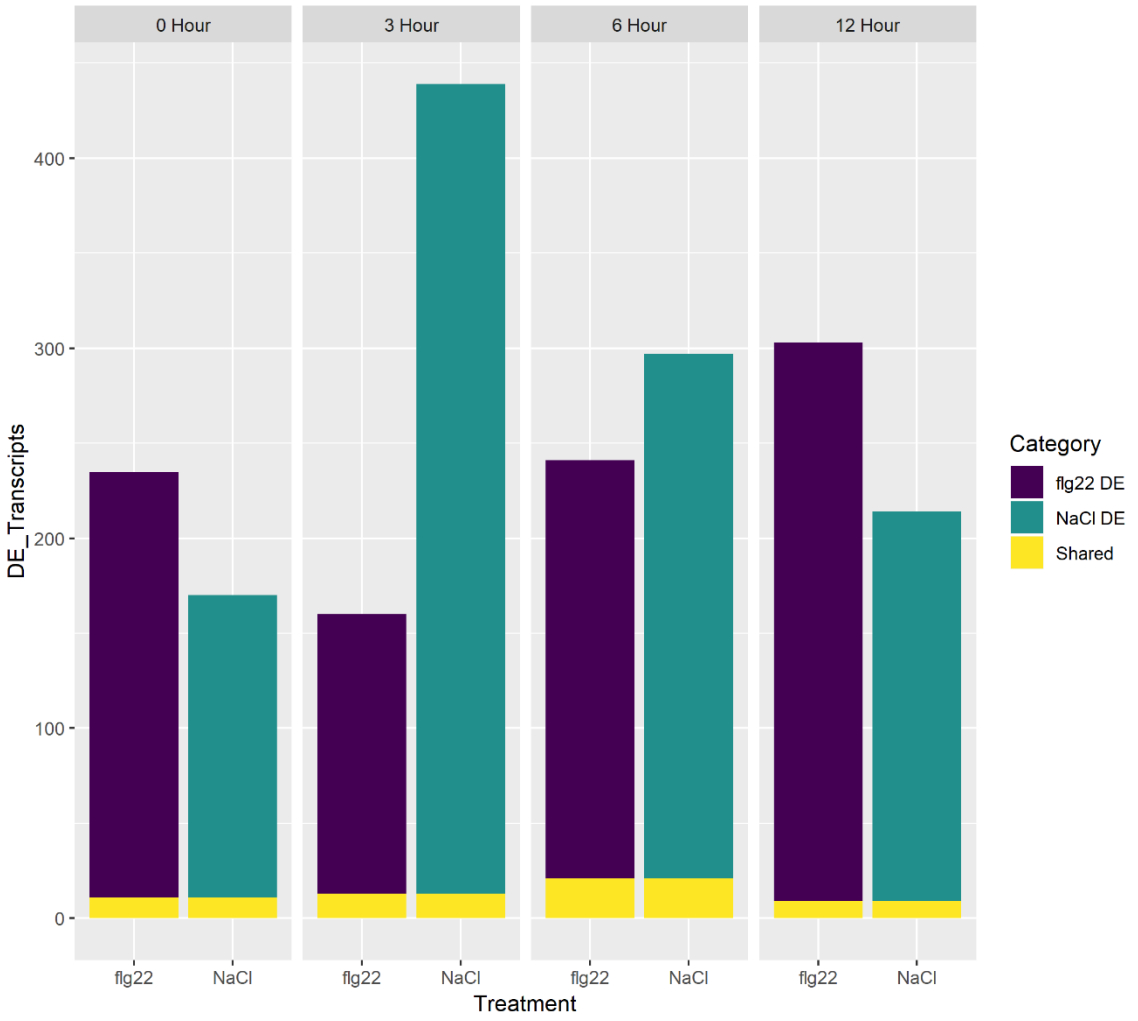


Figure D.3 Time-series comparison of differentially expressed transcripts during NaCl and flg22 treatments

Differentially expressed transcripts were identified in both examined treatments. Within individual time points, differentially expressed transcripts shared between treatment types are highlighted in yellow.

Marker genes, previously found to be differentially expressed alongside *ILK1* during either flg22 or NaCl treatment showed only slight changes in expression. These differences were generally not considered significant under the current significance thresholds. An exception is the *WRKY29* gene, which was found to be upregulated 70% following flg22 treatment after correction for confounding variables associated with time; previous examinations of *WRKY29* in *ILK1* knockdown mutants have reported similar increases in expression (9). FRK1 was another marker gene found to previously be upregulated in *ILK1* knockdown plants in response to flg22. The current dataset suggests only a slight 17% increase in expression. The salt stress marker *Rd29A* was also previously reported as showing significantly lower expression in *ILK1* knockdown mutants during salt stress responses. Correction for confounding variables associated with time did not produce a significant change in *Rd29A* expression following NaCl treatment in *ILK1* knockdown mutants.

Conclusion

High-quality reads were assembled into genes and transcripts which mapped with high precision and sensitivity to the reference Arabidopsis genome, showing high differential expression along the abiotic and biotic stresses and reduced transcriptional reprogramming for the *ILK1* gene knockdown. Differentially expressed genes were not identified using multiple testing correction thresholds within either treatment or at any examined time point. A possible explanation for this observation might originate in the function of *ILK1*. *ILK1* is a Raf-like MAP3K protein kinase functioning in signal transduction pathways. Alternative signaling kinases or different pathways may work alongside *ILK1* to respond to perceived stresses, resulting in functional redundancy where

the measurable differences in gene expression are muffled by alternative genes functioning to activate similar sets of transcripts for identical downstream responses. However, subsets of differentially expressed genes – including a known immune response marker *WRKY29* – were identified when eliminating the multiple testing correction, indicating that stress related transcriptional effects are associated with *ILK1*-mediated pathways. Further, insights into the role of *ILK1* from these datasets might also be gained through the use of alternative differential expression analysis pipelines.